# Multi modal images analysis and processing in oncology

Mathieu Hatt

université de bretagne occidentale

## Faculté de médecine de Brest

# Habilitation à diriger les recherches (HDR)

**"Analyse et traitement d'images multi modales en oncologie"**

# Habilitation-Accreditation to supervise research

**"Multi modal images analysis and processing in oncology"**


**Mathieu Hatt, PhD**

Laboratoire de Traitement de l'Information Médicale

INSERM U1101 Brest - France



Jury

Rapporteur : Pierre Celsis, DR INSERM, Montpellier
Rapporteur : Roland Hustinx, PU-PH, Liège (Belgique)
Rapporteur : Vincent Rodin, PU, Brest
Examinateur : Dimitris Visvikis, DR INSERM, Brest
Examinatrice : Catherine Cheze-Le Rest, MCU-PH, Brest
Examinateur : Fabien Salzenstein, MCU, Strasbourg

**2012**

# Outline

## I - Curriculum Vitae

1) Civil status

| | |
|---|---|
| Name: | **Hatt Mathieu** |
| Birth date and place: | **October the 5th, 1981 in Strasbourg, France** |
| Nationality: | **French** |
| Marital status: | **Single** |

2) Education (France)

- **2008**: **PhD** from the University of Brest, with highest honors.
  « Automatic determination of functional volumes in emission imaging for oncology applications » -1st prize of IEEE France for best biomedical thesis in 2008.

- **2004**: **Master in computer sciences** from the University of Strasbourg, with honors.
  Options: image processing, artificial intelligence, parallel computing, networks, algorithmic for graphics, bioinformatics.

- **2002**: **Licence in computer sciences** from the University of Strasbourg.

- **1999**: **Scientific baccalauréat**, specialization in physics and chemistry, with honors.

3) Research formation

- 11/2011-present **Stichting Maastricht Radiation Oncology (MAASTRO)**, Maastricht, the Netherlands. Director: Prof. P. Lambin. Supervisor : Prof. P. Lambin.

- 2005-2011 **Laboratory of medical information processing (LaTIM INSERM U650)**, Brest. Director: Pr. C. Roux. Supervisor: DR D. Visvikis.

- 2004: 6 months trainee in the **laboratory of sciences of image, computer sciences and remote sensing (LSITT UMR 7005)**, Strasbourg.
  Supervisors: Prof. C. Collet and F. Salzenstein.

4) Additional formation

- **2006**: 7th IEEE EMBS International Summer School on Biomedical Imaging, Berder Island, France.
- **2010**: visiting fellow in MRC clinical center, Hammersmith Hospital, London, UK.

## II - Professional experience

2002-2003: Trainee (two voluntary training periods of 2 months from July to August) as a computer scientist in training, in the astronomical observatory of Strasbourg (UMR 7550), stellar datacenter of Strasbourg (CDS).

- In charge of JAVA developments for SIMBAD stellar catalogs interface under the supervision of M. Wenger.

2003-2004: Trainee in the laboratory of sciences of image, computer sciences and remote sensing (LSITT UMR 7005) for 1st and 2nd year 6 months projects of master in computer sciences.

- Team « automation, vision and robotics » for the design of an expert system for robot control under the supervision of S. Besse.
- Team « models, images and vision » for the development of fuzzy hidden Markov chains for astronomical images segmentation under the supervision of C. Collet and F. Salzenstein.

2005-2008: PhD student in the laboratory of medical information processing (LaTIM INSERM U650), team « Quantitative multi modality imaging for diagnosis and therapy » under the supervision of D. Visvikis and C. Roux

- Development of methodologies dedicated to metabolically active tumor volume delineation in PET images for oncology applications.

2009-2011: Post-doc fellow in the laboratory of medical information processing (LaTIM INSERM U650), team « Quantitative multi modality imaging for diagnosis and therapy ».

- Principal investigator on ANR project SIFR[1] and in charge of supervising trainees and PhD students in the team.

2011-2012: Research fellow in the imaging and radiotherapy research department, MAASTRO lab, Maastricht, the Netherlands.

- In charge of collaborative research projects regarding the use of PET/CT imaging for the prediction of therapy response and prognosis in radiotherapy. Co-supervising two PhD students.

[1] Segmentation of functional images for radiotherapy, ANR TEC 2008, 500k€ (250k€ funded by ANR).

## III - Educational activities

1) Supervision and co-supervision of trainees and PhD students in the LaTIM

PhD students (supervision)

- 2009-2011: Simon David (thesis viva 13/12/2011)
  Image analysis for therapy response studies in PET
- 2009-2011: Amandine Le Maître (thesis viva 1st semester of 2012)
  Realistic simulations, automatic segmentation and dosimetry in PET/CT imaging
- 2011-2013: Houda Hanzouli (started October 2011)
  Multi resolution image analysis for multi modal imaging

PhD students (co-supervision)

- 2008-10: Adrien Le Pogam (thesis viva 04/2010)
  Partial volume effects correction in emission imaging
- 2010-12: Florent Tixier (thesis viva 2012)
  Characterization of tracer uptake heterogeneity in PET using textural features

Trainees (supervision)

- 2011: M. Sayed (ISEN engineer, 3 months) (collaboration with INSERM U613, Brest)
  Automatic registration of fluorescence images of mice for gene transfer applications
- 2010: T. Merlin (ISEN engineer, 6 months)
  Development and validation of an automatic algorithm for estimation and comparison of PET delineations for oncology

Trainees (co-supervision)

- 2011: Hela Rezgui (ENSI engineer, 9 months)
  Multivariate analysis for prognosis and response prediction in esophageal and head & neck cancers
- 2011: Houda Hanzouli (ENSI engineer, 6 months)
  PET images denoising using combined wavelet and curvelet transforms

2) Teaching (University of Brest)

Since 2010: DES of radiotherapy
PET Physics and use of PET imaging in radiotherapy applications
15h / year

Since 2009: master 2 SIBM (Signal and Image in Biology and Medicine)
PET/CT imaging, digital medical image processing, PET quantification
9h / year
Advanced image segmentation techniques and applications in medical imaging
3h / year

# IV - Publications and communications

## 1) Original articles in peer-reviewed journals

1. A. Le Maitre, D. Visvikis, C. Cheze-Le Rest, O. Pradier, **M. Hatt**. Dose prescription adapted to functional tumor $^{18}$F-FDG heterogeneities: the influence of contrast and size of sub-volumes. Physics in Medicine and Biology 2012; in revision

2. S. David, D. Visvikis, Q. Quellec, P. Fernandez, M. Allard, C. Roux, **M. Hatt**. Image change detection using paradoxical theory for patient follow-up quantitation and therapy assessment. IEEE Transactions on Medical Imaging 2012; in revision.

3. F. Tixier, **M. Hatt**, C. Cheze Le Rest, A. Le Pogam, L. Corcos, D. Visvikis. Reproducibility of tumor uptake heterogeneity characterization through textural feature analysis in $^{18}$F-FDG PET imaging. Journal of Nuclear Medicine 2012; in press.

4. **M. Hatt**, A. Le Pogam, D. Visvikis, O. Pradier, C. Cheze le Rest. Impact of partial volume effects correction on the predictive and prognostic value of baseline $^{18}$F-FDG PET images in esophageal cancer. Journal of Nuclear Medicine 2012;53(1):12-20.

5. **M. Hatt**, C. Cheze le Rest, A. van Baardwijk, P. Lambin, O. Pradier, D. Visvikis. Impact of tumor size and tracer uptake heterogeneity in $^{18}$F-FDG PET and CT Non–Small Cell Lung Cancer tumor delineation. Journal of Nuclear Medicine 2011;52(11):1690-7.

6. S. David, D. Visvikis, C. Roux, **M. Hatt**. Multi observation PET image analysis for patient follow-up quantitation and therapy assessment. Physics in Medicine and Biology 2011;56(18):5771-88. [Featured free article as *editor's choice]*

7. A. Le Pogam, **M. Hatt**, P. Descourt, N. Boussion, C. Tsoumpas, FE. Turkheimer, C. Prunier-Aesch, J-L. Baulieu, D. Guilloteau, D. Visvikis. Evaluation of a 3D local multi-resolution algorithm for the correction of partial volume effects in positron emission tomography. Medical Physics 2011;38(9):4920-4933. [Cover of the issue]

8. **M. Hatt**, D. Visvikis, O. Pradier, C. Cheze-le Rest. Baseline $^{18}$F-FDG PET image derived parameters for therapy response prediction in œsophageal cancer. European Journal of Nuclear Medicine and Molecular Imaging 2011;38(9):1595-1606.

9. **M. Hatt**, D. Visvikis, N. M. Albarghach, F. Tixier, O. Pradier, C. Cheze-le Rest. Prognostic value of $^{18}$F-FDG PET image-based parameters in œsophageal cancer and impact of tumor delineation methodology. European Journal of Nuclear Medicine and Molecular Imaging 2011;38(7):1191-1202.

10. F. Tixier, C. Cheze Le Rest, **M. Hatt**, N. M. Albarghach, O. Pradier, J-P. Metges, L. Corcos, D. Visvikis. Intra-tumor heterogeneity characterized by textural features on baseline $^{18}$F-FDG PET images predicts response to concomitant radio-chemotherapy in esophageal cancer. Journal of Nuclear Medicine 2011;52(3):369-378.

11. **M. Hatt**, C. Cheze le Rest, N. M. Albarghach, O. Pradier, D. Visvikis. Robustness and repeatability of image segmentation approaches dedicated to PET tumor uptake volume delineation. European Journal of Nuclear Medicine and Molecular Imaging 2011;38(4):663-672.

12. **M. Hatt**, C. Cheze-Le Rest, E. O. Aboagye, L. M. Kenny, L. Rosso, F. E. Turkheimer, N. M. Albarghach, O. Pradier, D. Visvikis. Reproducibility of $^{18}$F-FDG and 3'-Deoxy-3'-$^{18}$F-Fluorothymidine PET Tumor Volume Measurements. Journal of Nuclear Medicine 2010;51(9):1368-1376.

13. **M. Hatt**, C. Cheze le Rest, P. Descourt, A. Dekker, D. De Ruysscher, M. Oellers, P. Lambin, O. Pradier, D. Visvikis. Accurate automatic delineation of heterogeneous functional volumes in positron emission tomography for oncology applications. International Journal of Radiation Oncology Biology Physics 2010;77(1):301-308.

14. A. Le Maitre, W.P. Segars, S. Marache, A. Reilhac, **M. Hatt**, S. Tomei, C. Lartizien, D. Visvikis. Incorporating patient specific variability in the simulation of realistic whole body $^{18}$F-FDG distributions for oncology applications. Proceedings of the IEEE Special Issue on Computational anthropomorphic anatomical models, 2009;97(12):2026-2038.

15. **M. Hatt**, C. Cheze le Rest, A. Dekker, D. De Ruysscher, M. Oellers, P. Lambin, C. Roux, D. Visvikis. Une nouvelle méthode de détermination automatique des volumes fonctionnels pour les applications de l'imagerie d'émission en oncologie. Ingénierie et Recherche BioMédicale (numéro spécial RITS 2009) 2009;34(4):144-149.

16. **M. Hatt**, A. Turzo, C. Roux, D. Visvikis. A fuzzy locally adaptive Bayesian segmentation approach for volume determination in PET. IEEE Transactions on Medical Imaging 2009;28(6):881-893.

17. N. Boussion, C. Cheze Le Rest, **M. Hatt**, D. Visvikis. Incorporation of wavelet based denoising in iterative deconvolution for partial volume correction in whole body PET imaging. European Journal of Nuclear Medicine and Molecular Imaging 2008;36(7):1064-75.

18. N. Boussion, **M. Hatt**, F. Lamare, C. Cheze Le Rest, D. Visvikis. Contrast enhancement in emission tomography by way of synergistic PET/CT image combination. Computer Methods and Programs in Biomedicine 2008;90(3):191-201.

19. **M. Hatt**, F. Lamare, N. Boussion, A. Turzo, C. Collet, F. Salzenstein, C. Roux, K. Carson, P. Jarritt, C. Cheze-Le Rest, D. Visvikis. Fuzzy hidden Markov chains segmentation for volume determination and quantitation in PET. Physics in Medicine and Biology 2007;52(12):3467-3491.

20. F. Salzenstein, C. Collet, S. Lecam, **M. Hatt**. Non-stationary fuzzy Markov chain. Pattern Recognition Letters 2007;28(16):2201-2208.

21. N. Boussion, **M. Hatt**, F. Lamare, Y. Bizais, A. Turzo, C. Cheze-Le Rest, D. Visvikis. A multi resolution image based approach for correction of partial volume effects in emission tomography. Physics in Medicine and Biology 2006;51(7):1857-18766.

2) Reviews in peer-reviewed journals

**M. Hatt**, N. Boussion, C. Cheze-le Rest, D. Visvikis, O. Pradier. Metabolically active volumes automatic delineation methodologies in PET imaging: review and perspectives. Cancer/Radiothérapie 2011; online in october.

3) Letters to the editor of peer-reviewed journals

**M. Hatt**, D. Visvikis, C. Cheze Le Rest. Regarding Autocontouring and Manual Contouring: Which Is the Better Method for Target Delineation Using $^{18}$F-FDG PET/CT in Non–Small Cell Lung Cancer? By K. Wu et al. Journal of Nuclear Medicine 2011;52(4):658.

**M. Hatt**, D. Visvikis. Defining radiotherapy target volumes using $^{18}$F-Fluoro-Deoxy-Glucose Positron Emission Tomography: still a Pandora box?: in regard to Devic et al. International Journal of Radiation Oncology Biology Physics 2010; 78(5):1605.

4) Book chapters

**M. Hatt**, D. Visvikis, Chapitre 4 : Tomographie par émission de positons et tomographie d'émission monophotonique dynamiques, in 'Imagerie Dynamique Cardiaque : Systèmes et Techniques d'acquisition', P. Clarysse & F. Frouin publishing, 2011.

5) Communications and invited talks in peer-reviewed international conferences

See appendix for full list.

## V - Scientific societies, grants, industrial and research partnerships, awards

### Scientific societies and professional committees

- Member of the following scientific societies

  - IEEE (Institute of Electrical and Electronics Engineers)
  - AAPM (American Association of Physicists in Medicine)
  - SNM (Society of Nuclear Medicine)
  - SFGBM (French Society of BioMedical enGineering)

- Member of AAPM Taskgroup n° 211 [2]
  Classification, Advantages and Limitations of the Auto-Segmentation Approaches for PET

- Member of SNM computer & instrumentation council

- Substitute member for France of COST European action TD-10-07 [3]
  Bimodal molecular imaging technologies coupling PET and MRI for in vivo visualization of pathologies and biological processes

- Associate editorial board member of the American Journal of Nuclear Medicine and Molecular Imaging[4]

- Referee for the following journals:

  - European Journal of Nuclear Medicine
  - IEEE Transactions on Medical Imaging
  - IEEE Transactions on Information Technology in BioMedicine
  - IEEE Transactions on Nuclear Science
  - Journal of Applied Clinical Medical Physics
  - Journal of Nuclear Medicine
  - Journal of Nuclear Medicine and Radiation Therapy
  - Medical Physics
  - Physics in Medicine and Biology

### Grants

- French national research agency (ANR, 250 k€)
- Ligue contre le cancer (30 k€)
- Institut Telecom (55 k€)
- PhD student grants (190 k€)

[2] http://www.aapm.org/org/structure/default.asp?committee_code=TG211

[3] http://www.cost.eu/domains_actions/mpns/Actions/TD1007?management

[4] http://www.ajnmmi.us/editorialboard.html

## Research partnerships

National

- CHRU Brest
- CHRU Bordeaux
- CHRU Toulouse

International

- MAASTRO lab – Maastricht (The Netherlands)
- Nijmegen (The Netherlands).
- CHU Liège (Belgium)
- MRC Clinical Sciences Centre Hammersmith (London, UK)
- UCLH (University College London Hospital) (London, UK)
- Royal Surrey County Hospital (Guilford, UK)
- University of Wisconsin, Madison (USA)
- University of Seattle, Washington (USA)
- University of Maryland, Baltimore (USA)
- MD Anderson, Houston, Texas (USA)
- University of Washington (Seattle, USA)
- Munich (Germany)
- Freiburg (Germany)

## Industrial partnerships

- Research agreements with Philips Healthcare and Siemens Healthcare regarding the exploitation of the PET segmentation FLAB algorithm

## Awards

- « Best-in-physics » paper in AAPM annual meeting 2011[5].
- « Young investigators award » of the New trends in molecular imaging and nuclear medicine conference, 2009.
- « 1st prize research » for best biomedical PhD thesis in 2008, from IEEE-France.
- Travel grant from NSS-MIC 2008 conference based on the « scientific excellence of the submitted contribution ».

## Other assignments in the laboratory

- In charge of the internal review committee for the articles written within the group

- Delegate for the post-doc fellows of the laboratory.
- In charge of the organization and planning for the monthly meeting of the laboratory.

[5] http://www.aapm.org/m/mtg/absdetail.asp?mid=59&sid=3844&aid=16323

Résumé en français (French summary)

Avec une formation initiale en sciences de l'informatique et une spécialisation image, mes activités de recherche actuelles concernent le traitement et l'analyse de l'information et de l'image pour des applications en médecine, plus particulièrement l'oncologie et la radiothérapie. Plus spécifiquement, je m'intéresse à la segmentation et la classification automatique pour la définition des contours d'organes et de tumeurs, au filtrage du bruit et à la déconvolution pour l'amélioration qualitative et quantitative, et plus récemment, aux modèles multi observation pour la prise en compte des images multi modales, et la fusion d'informations pour l'aide à la décision dans la prise en charge des patients. Je poursuis ces thématiques spécifiquement dans le cadre de l'utilisation de l'imagerie TEP/TDM (Tomographie par Emission de Positons et scanner X) en oncologie et radiothérapie.
Mes activités de recherche ont pris place dans le contexte de l'équipe « imagerie multi modale quantitative pour le diagnostic et la thérapie » du laboratoire INSERM U650 de traitement de l'information médicale (LaTIM). Ce contexte a garantit un travail d'équipe pluridisciplinaire, en collaboration notamment avec des radiothérapeutes, des médecins nucléaires, des physiciens, des ingénieurs, des mathématiciens et des informaticiens.
En tant que doctorant, ma principale contribution a été le développement d'une méthode originale de segmentation d'image adaptée à la définition des volumes fonctionnels des tumeurs sur les images TEP. Lors de mon post-doctorat, j'ai poursuivi la validation de la précision, de la robustesse et de la reproductibilité de cette approche dans le cadre d'un projet ANR pour lequel j'ai reçu un financement de deux ans et demi. J'ai également étudié au cours de ces deux dernières années l'impact d'une telle méthode dans de nombreuses applications, telles que la dosimétrie en planification de traitement en radiothérapie, et la prise en charge des patients en oncologie.
Au cours de ces six dernières années, j'ai été de plus en plus impliqué dans des travaux de recherche connexes menés par d'autres doctorants et post-doctorants. Ces travaux incluent la fusion d'images TEP pour le suivi temporel quantitatif, les simulations réalistes et l'évaluation dosimétrique, la caractérisation de l'hétérogénéité intra tumorale des traceurs TEP par analyse de texture, et la réduction des effets de volume partiel et du bruit en imagerie d'émission. J'ai assumé la responsabilité de co-encadrant de plusieurs stagiaires et doctorants de l'équipe sous la direction du directeur de recherche D. Visvikis. Cette responsabilité inclus des réunions hebdomadaires et des discussions régulières avec les étudiants, l'aide à la mise en place des expériences et protocoles de validation, à l'analyse des résultats, la correction de rapports de stage, d'articles et de manuscrits de thèse, ainsi que réfléchir à des solutions aux problèmes tant théoriques que techniques. Je travaille actuellement en tant que chercheur associé au département de recherche en imagerie et radiothérapie de Maastricht (MAASTRO) aux Pays-bas.

Au cours des prochaines années, mon projet de recherche sera dédié au développement d'un contexte flexible et robuste permettant la modélisation et l'analyse semi-automatique d'ensemble d'images médicales multi modales, multi résolutions et multi temporelles, telles que TEP/TDM, TEMP/TDM, TEP/IRM, multi IRM ou TEP avec différents traceurs, ainsi que des acquisitions dynamiques. Ce développement permettra de déduire de nouveaux modèles prédictifs et des outils de décision adaptés à diverses applications cliniques tels que les cancers de l'œsophage, rectal, pulmonaire ou ORL, par la fusion de toute l'information disponible (imagerie, génétique, phénotypes et rapports cliniques). Ce projet se construira en partie sur les travaux préliminaires réalisés avec des doctorants venant de soutenir et en passe de terminer leur thèse, et sur les thèses de deux nouvelles doctorantes que j'encadrerai à partir d'octobre 2011 et courant 2012, recrutées sur des financements que j'ai contribué à obtenir en 2010-2011.

# VI - Overview of past and present research activities

With an initial formation in theoretical computer sciences with a focus on image processing and analysis, my current research activities deal with image and information processing and analysis for applications in medicine, namely oncology and radiotherapy. More specifically, my research interests are image automatic segmentation and classification for organs and tumors delineation, image denoising and deconvolution for qualitative and quantitative improvement, and more recently, multi observation models for multi modal imaging and information fusion for computer-aided decision making in patients management. These developments are especially considered within the context of the use of Positron Emission Tomography and Computed Tomography (PET/CT) for oncology and radiotherapy applications.

My research activities have been and are still carried out within the team "quantitative multi modal imaging for diagnosis and therapy", in the LaTIM INSERM U1101. This framework ensures a multi disciplinary teamwork, in collaboration with radiation oncologists, nuclear medicine physicians, physicists, engineers, mathematicians and computer scientists.

As a PhD student, my main contribution to the field has been the development of image segmentation algorithms dedicated to the automated delineation of metabolically active tumor volumes in PET images, with a specific focus on adapting the methodology to specific characteristics of the processed images. As a post-doctoral fellow, I have been further investigating the accuracy, robustness and reproducibility of this methodology within a project for which I had obtained funding from the French research agency (ANR) for two and a half years. I have also been investigating the impact of such methodology and its resulting tumor volumes measurements in various applications such as the dosimetry impact in radiotherapy treatment planning or patient management and therapy assessment in oncology.

During the last six years I have also been more and more involved in research developments by several PhD students and post-doctoral fellows, such as PET images fusion for quantitative follow up, realistic simulations and dosimetry evaluation, methodologies for reduction of partial volume effects and noise in emission imaging, and textural features analysis for characterization of tracer uptake heterogeneity within tumors. Overall, I have been acting as co-supervisor of most of the trainees and PhD students of the team under the direction of Research Director D. Visvikis. This responsibility included weekly meetings and discussions with the students, help with designing experiments and analyzing the results, writing of thesis reports and research articles, as well as finding solutions to theoretical problems and technical issues.

In the next few years, my research project will be focused on the development of a robust and flexible framework for the modeling and the automatic analysis of multi modality, multi resolution, multi observation and multi temporal images datasets, such as PET/CT, SPECT/CT, PET/MRI, multi MRI or PET tracers imaging, as well as dynamic acquisitions. This development will allow deriving new predictive models and decision tools by fusion of the available multi source information (imaging, genetics and other clinical data), validated in various models such as esophageal, rectal, lung, or head & neck cancers. Additional applications in neurology (for example Alzheimer's disease) might also be explored. This project will be based on previous developments by PhD students that are almost finished with their thesis, and will focus on new PhD students beginning their work under my supervision in October 2011 and early 2012, whose recruitment was possible thanks to grants I contributed to obtain in 2010-2011.

# 1. Research during the PhD

## A. PhD methodological developments

One of the main factors of error for semi-quantitative analysis in positron emission tomography (PET) imaging for diagnosis and patient follow up (1), as well as new flourishing applications like image guided radiotherapy (2), is the methodology used to define the volumes of interest in the functional images. This is explained by poor image quality in emission tomography resulting from noise and partial volume effects (3) induced blurring, as well as the variability of acquisition protocols, scanner models and image reconstruction procedures (4). Manual delineation of the metabolically active tumor volumes (MATV) is extremely subjective and suffers from major inter and intra observer variability (5). In addition, it is especially tedious and time consuming; therefore it is never used in clinical practice. The majority of previously published approaches were based at the time (before 2005) on deterministic binary thresholding (6) that are not robust to contrast variation and noise (7). In addition, these methodologies are unable to correctly handle heterogeneous uptake inside tumors (8). The objective of my thesis was to develop an automatic, robust, accurate and reproducible 3D image segmentation approach for the functional volumes determination of tumors of all sizes and shapes, and whose activity distribution may be strongly heterogeneous. The approach I have developed is based on a statistical image segmentation framework, combined with a fuzzy measure, which allows to take into account both noisy and blurry properties of emission images (9). A first development was carried out using fuzzy hidden Markov chains as a spatial model [1], which gave satisfying results except for small structures (either small lesions or small sub volumes within a lesion, as well as complex shapes and contours) [2]. A second development was carried out to solve these issues and was named FLAB for Fuzzy Locally Adaptive Bayesian [3]. It still exploited a stochastic iterative parameters estimation and a fuzzy measure as in the first method, however the hidden Markov chains model was replaced by a locally adaptive model of the voxel and its neighbors for the estimation and segmentation. This method was also improved in order to be able to consider up to three classes in the images, in order to account for heterogeneous activity (either in the background or within the tumor). FLAB was evaluated using a large array of datasets, comprising both simulated and real acquisitions of phantoms and tumors. The results obtained on phantom acquisitions allowed validating the accuracy of the segmentation with respect to the size of considered structures, down to 13-17 mm in diameter as well as its robustness with respect to noise, contrast variation, and acquisition parameters. The performance of the developed algorithm was shown to be superior to threshold-based methodologies and other clustering algorithms. The results demonstrated the ability of the developed approach to accurately delineate tumors with complex shapes and activity distributions, as illustrated in figure 1, in which the result of FLAB is compared to the results on an adaptive threshold approach used by two different clinicians on patient image with an esophageal lesion. This illustrates well the ability of FLAB to obtain a complete tumor volume in case of heterogeneous activity, contrary to the adaptive threshold approach. It also emphasizes the lack of repeatability of the adaptive threshold method due to heterogeneous background uptake in the mediastinum.

The FLAB algorithm was also able to delineate multiples regions inside the tumor [4]. Some of the datasets used for the accuracy evaluation were generated using realistic Monte Carlo simulations that were improved in several ways in collaboration with another PhD student (A. Le Maitre), including patient-specific anatomical properties and complex non spherical shaped tumors exhibiting heterogeneous tracer uptake [5]. Both robustness and accuracy results demonstrated that the proposed methodology may be used in clinical context for diagnosis and patients follow up, as well as for radiotherapy treatment planning and "dose painting", facilitating optimized dosimetry and potentially reduced doses delivered to healthy tissues around the tumor and nearby organs.

**[1] F. Salzenstein, C. Collet, S. Lecam, M. Hatt, Non-stationary fuzzy Markov chain, PRL 2007**
**[2] Hatt et al. Fuzzy hidden Markov chains segmentation for volume determination and quantitation in PET, PMB 2007**
**[3] Hatt et al. A fuzzy locally adaptive Bayesian segmentation approach for volume determination in PET, IEEE TMI 2009**
**[4] Hatt et al. Accurate automatic delineation of heterogeneous functional volumes in positron emission tomography for oncology applications, IJROBP 2010**
**[5] A. Le Maitre, W.P. Segars, S. Marache, A. Reilhac, M. Hatt, et al. Incorporating patient specific variability in the simulation of realistic whole body $^{18}$F-FDG distributions for oncology applications, Proceedings of the IEEE 2009**

This work was recognized by an award co-delivered by the French section of the IEEE and the French society of biomedical engineering (SFGBM) for the best research PhD thesis in biomedical imaging defended in 2008. A conference paper describing the latest developments and validation of the FLAB approach was recognized in 2008 by the IEEE-Medical Imaging Conference committee through a trainee grant based on the scientific excellence of the contribution. According to Google Scholar, the four methodological papers [1-4] add up to more than 100 citations, almost 50 of them for the FLAB paper [3].

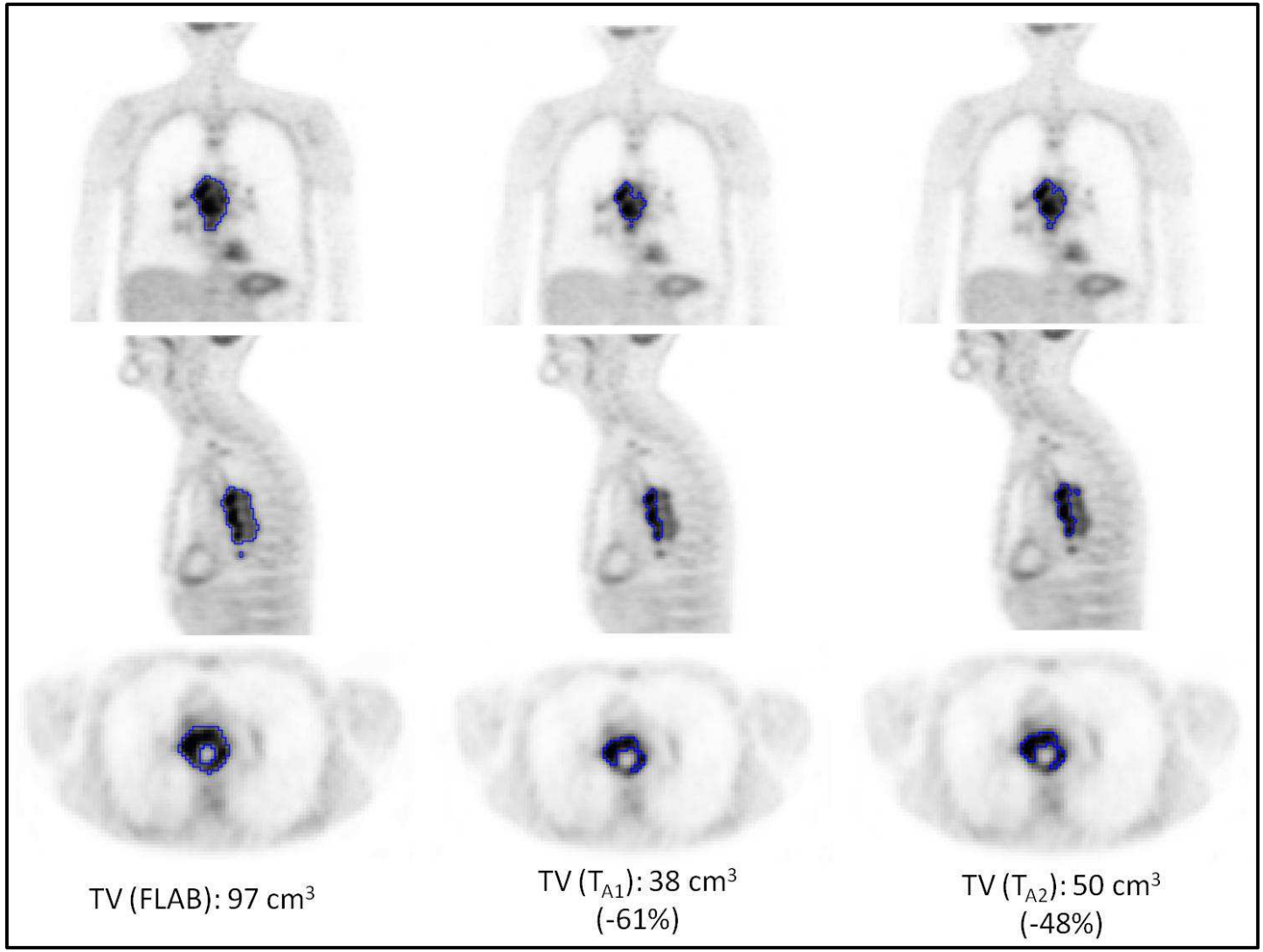



**Fig.1 Coronal, sagital and axia views of a 18F-FDG PET image of a patient with esophageal cancer (large heterogeneous MATV). Delineation (blue contours) using FLAB (on the left) and adaptive threshold with two different observers (on the right and in the middle). Note the significant underestimation obtained with adaptive threshold (both users) due to the heterogeneity.**

## B. Additional methodological developments

There are numerous other pitfalls and sources of errors in PET imaging. Most importantly, the level of noise and its characteristics, as well as partial volume effects (PVE), lead to significant quantitative biases and qualitatively degraded images (4). Both aspects are also closely related to the problem of automatic delineation of PET images. Therefore, during my PhD I have been working closely with a post-doctoral fellow (N. Boussion) and a fellow PhD student in the team (A. Le Pogam), on denoising and deconvolution methodologies dedicated to emission imaging. I have contributed to several developments, including two PVE correction methods and a denoising approach. The first methodology was based on the use of anatomical high resolution details from the co-registered morphological images (such as CT or MRI) in order to correct for the spill in and spill out effects of partial volume effects in the associated functional dataset. This Mutual Multi resolution Analysis (MMA) method exploited spatial-frequency analysis, namely wavelet transforms, and extracted structural details from these decompositions in order to derive a model linking both image modalities [1-2]. Another approach was also developed in order to correct for PVE in emission images without requiring associated high resolution anatomical datasets, or in cases (and there are many) for which no spatial correlation between the morphological and functional datasets can be exploited. This second approach was based on iterative deconvolution (10-11), which is a well known method for improving spatial resolution of images. However, such algorithms are associated with increased noise levels in the deconvolved images, which is not compatible with their subsequent clinical use. Therefore, we designed a denoising methodology dedicated to emission imaging, based on the filtering of wavelet coefficients using a Bayesian based method to discriminate between noise and information in the wavelet domain (12). This methodology was then included in the iterative deconvolution process in order to control the noise propagation additively introduced by each iteration of the deconvolution. This allowed significantly reducing the blur in emission imaging without significant addition of noise. The validation of the method demonstrated its ability to restore accurate quantitative measurements in the images, while providing full whole-body corrected images [3]. Note that the developed denoising method was also considered as a standalone denoising approach for emission imaging (see part 2.3)

**[1] N. Boussion, M. Hatt, et al. A multiresolution image based approach for correction of partial volume effects in emission tomography, PMB 2006**

**[2] N. Boussion, M. Hatt, et al. Contrast enhancement in emission tomography by way of synergistic PET/CT image combination, CMBP 2008**

**[3] N. Boussion, C. Cheze Le Rest, M. Hatt, et al. Incorporation of wavelet based denoising in iterative deconvolution for partial volume correction in whole body PET imaging, EJNM 2008**

## 2. Post-doctoral work

During the last few months of my PhD, I applied for a research grant to the French National Research Agency (ANR Emergence TEC 2008 call to projects) and received a 250k€ grant for two and a half year for a project named SIFR (segmentation of functional images for radiotherapy, complete cost 500k€).
The goals of this project were to i) further validate the automatic delineation algorithm proposed during my PhD and ii) investigate its impact and value in the clinical setting through various studies.
The following describes my main research activities within this SIFR project, which included supervision of trainees and PhD students, as well as other contributions to additional methodological developments carried out in the team by various students under my co-supervision.

### A. Robustness, repeatability and reproducibility of MATV measurements in PET imaging

Automatic delineation approaches for MATV measurements in PET images may be of interest for applications such as target volume definition in radiotherapy for scenarios of dose redistribution, boosting or painting (13), and oncology applications such as diagnosis, prognosis and prediction or assessment of response to therapy (early or late during treatment) (1, 14). Their use however faces several pitfalls.

First, there is a clear lack of standardization of acquisition and reconstruction protocols across clinical centers (4, 15). Each one has its own scanner model and associated vendor-specific iterative reconstruction algorithm (and associated pre or post filtering options and voxel sizes for the reconstruction grid), with a specific set of chosen and often optimized parameters for their specific acquisition protocols (injected dose, time between injection and acquisition, acquisition duration, etc). Such differences lead to images that have vastly different properties of noise levels, signal-to-noise ratios, textures, and spatial resolution. The robustness of the method used to delineate MATVs on images from different centers is therefore crucial, especially when considering multi-centric clinical studies. One objective was therefore to investigate the robustness of existing methods.
Second, the reproducibility of PET scans is known to be limited, even with highly standardized acquisition and reconstruction protocols. The reproducibility of maximum activity measured in tumors had been previously assessed by various authors using double baseline PET scans (carried out at a few days interval with a procedure as identical as possible), and reported upper and lower reproducibility limits of about ±15 to 30% (16-17). These results constituted the basis for the definition of confidence intervals regarding the required maximum activity variation between pre and post treatment PET scans to characterize patient responding, stable or progressive disease in solid tumors (1). If the MATVs and their associated measurements (mean SUV for example) are to be used within the same context, it is mandatory to evaluate their reproducibility on double baseline scans, which was a second objective.

Third, the repeatability is an important characteristic of any automated approach, since one of the major advantages of using automatic algorithms is the low inter and intra user variability, with respect to manual delineation, in addition to the gain in time. The evaluation of the repeatability was therefore a third objective.

a) Robustness

We designed the following study for assessing the robustness of the FLAB methodology with respect to other existing approaches (fixed and adaptive threshold). We considered a single physical phantom containing spheres of various diameters that can be filled with activity, as well as the background, in order to create a contrast between the sphere (simulating a simple tumor) and the background (simulating an homogeneous physiological background). Clearly, such homogeneous spheres on homogeneous background are insufficient to properly validate the accuracy of a delineation algorithm since tumors are often non spherical and exhibit heterogeneous tracer uptake. However, this is a proper tool to investigate robustness of the delineation with respect to varying acquisition conditions. The originality of this study was to consider acquisitions carried out on four state-of-the-art PET/CT scanners by all three vendors (Siemens, GE, Philips), including a time-of-flight model (Philips), and their associated reconstruction algorithms (OSEM, RAMLA and MLEM) and post-filtering options. In addition, a protocol was designed such as for each scanner model, different acquisition parameters would be available: two contrasts between the spheres and the background (around 4-5 to 1, and 8-10 to 1), three acquisition durations (1, 2 and 5 minutes) to investigate the noise, and two voxel sizes used in the reconstruction (2 or 4-5 mm in each dimension). This allowed for a wide range of image qualities, texture and properties to investigate the robustness. FLAB demonstrated significantly higher robustness (lower variability of the results) to varying acquisition and reconstruction parameters or across scanner models than the other methods [1].

b) Reproducibility

We designed the following study to investigate reproducibility of tumor volume measurements in PET images: two clinical datasets were considered, including esophageal lesions with $^{18}$F-FDG, and breast cancer with $^{18}$F-FLT, both with double baseline PET scans. At this occasion, I stayed as a visiting fellow in the MRC clinical center of the Hammersmith Hospital, in London, for the specific investigation of the reproducibility of tumor volumes measurements in breast cancer imaged with $^{18}$F-FLT. We demonstrated that using FLAB, tumor volumes could be automatically delineated on both FDG and FLT (despite the increased noise and lower contrast in FLT images) with reproducibility similar to the extraction of maximum activity with upper and lower limits of about 30%. On the contrary, using threshold-based methods, upper and lower limits were significantly higher (40-90%, consistent with another study (18)) and therefore less compatible for response monitoring purposes, especially early during treatment [2].

c) Repeatability

Repeatability was investigated for FLAB and other methods in both the previous studies considering different simulated and real clinical datasets, demonstrating much higher

repeatability with the automated methods (less than a few % variability) than manual delineation (15-30% inter and intra observer variability) [1-2].

**[1] M. Hatt, et al. Robustness and repeatability of image segmentation approaches dedicated to PET tumor uptake volume delineation, EJNM 2011.**
**[2] M. Hatt, et al. Reproducibility of $^{18}$F-FDG and 3'-Deoxy-3'-$^{18}$F-Fluorothymidine PET Tumor Volume Measurements, JNM 2010.**

This work was recognized by a "young investigator award" delivered by the conference "New trends in molecular imaging and nuclear medicine" in 2009. Despite being relatively recent papers, these two papers add up to 24 citations according to Google Scholar.

## B. Clinical value and impact of automatic MATV delineation in oncology and radiotherapy

Automatic delineation approaches for MATV measurements in PET images may be useful for several applications, for instance patient management and therapy monitoring in oncology, as well as tumor targeting in radiotherapy planning, dose redistribution and boosting.

a) Prognosis and response to therapy prediction

$^{18}$F-FDG PET has been identified as a powerful tool for diagnosis and prognosis in several cancer models, such as lung, esophageal, rectal and head and neck (19). In addition, the use of PET has been suggested to assess treatment response after the end of treatment or earlier during treatment (1). Finally, it has been suggested that it may be possible to predict response based on the baseline scan before treatment, which could improve patients' management. Potential non responders could indeed be identified before treatment, thus avoiding unnecessary toxicities. However, most of the studies have investigated the prognostic or predictive value of PET images by extracting maximum activity ($SUV_{max}$) only. In order to investigate the potential clinical value of accurate tumor volume delineation and the extraction of associated parameters from baseline $^{18}$F-FDG PET scans, we carried out a retrospective study on 50 patients treated for locally advanced esophageal cancer with concomitant radiochemotherapy. The results demonstrated that whereas standard clinical parameters and usual SUV measurements were neither prognostic nor predictive factors, measurements derived from metabolic volume were highly correlated with overall survival [1] or response to therapy [2], larger and more active volumes being associated with poor outcome and worse response. This is illustrated in figure 2. Receiver operating characteristic (ROC) curves are provided for the identification of complete responders or non responders, using various image-derived indices. MATV derived indices have significantly higher area under the curve (AUC) than SUV measurements.

In both these studies we also demonstrated that more accurate prediction could be achieved with more accurate methods. I also took part in the investigation of additional PET derived indices by co-supervising a PhD student (F. Tixier) working on the spatial patterns characterization of the tracer uptake heterogeneity within the tumor in PET images. Tumor heterogeneity has been identified as a potential factor of failed treatment and its characterization is therefore of potentially high clinical value (20). After accurate tumor

volume delineation, several parameters derived from textural features analysis of the voxels within the tumor could provide characterization and quantification of local and regional heterogeneity patterns. Several of these parameters were significantly correlated with response, higher heterogeneity being associated with poorer or lack of response [3]. Therefore, we demonstrated that it may be possible to improve patient management by identifying potential non responders before even initiating treatment by exploiting more fully the information contained in the baseline PET images. However, such a more complete analysis requires validated and robust semi-automated tools (21). Similarly to MATV measurements, evaluation of the specific reproducibility of these new indices is crucial to identify which parameters could be used for heterogeneity characterization during treatment for response monitoring. Such a study has been conducted and allowed identifying several local and regional heterogeneity parameters with sufficient reproducibility, but also excluding some of them that were characterized by very high variability across double baseline scans [4].

**[1] M. Hatt, et al. Prognostic value of $^{18}$F-FDG PET image-based parameters in œsophageal cancer and impact of tumor delineation methodology, EJNM 2011**

**[2] M. Hatt, et al. Baseline $^{18}$F-FDG PET image derived parameters for therapy response prediction in œsophageal cancer, EJNM 2011**

**[3] F. Tixier, C. Cheze Le Rest, M. Hatt, et al. Intra-tumor heterogeneity characterized by textural features on baseline $^{18}$F-FDG PET images predicts response to concomitant radio-chemotherapy in esophageal cancer, JNM 2011.**

**[4] F. Tixier, M. Hatt, et al. Reproducibility of tumor uptake heterogeneity characterization through textural feature analysis in $^{18}$F-FDG PET imaging, JNM 2012; in press.**

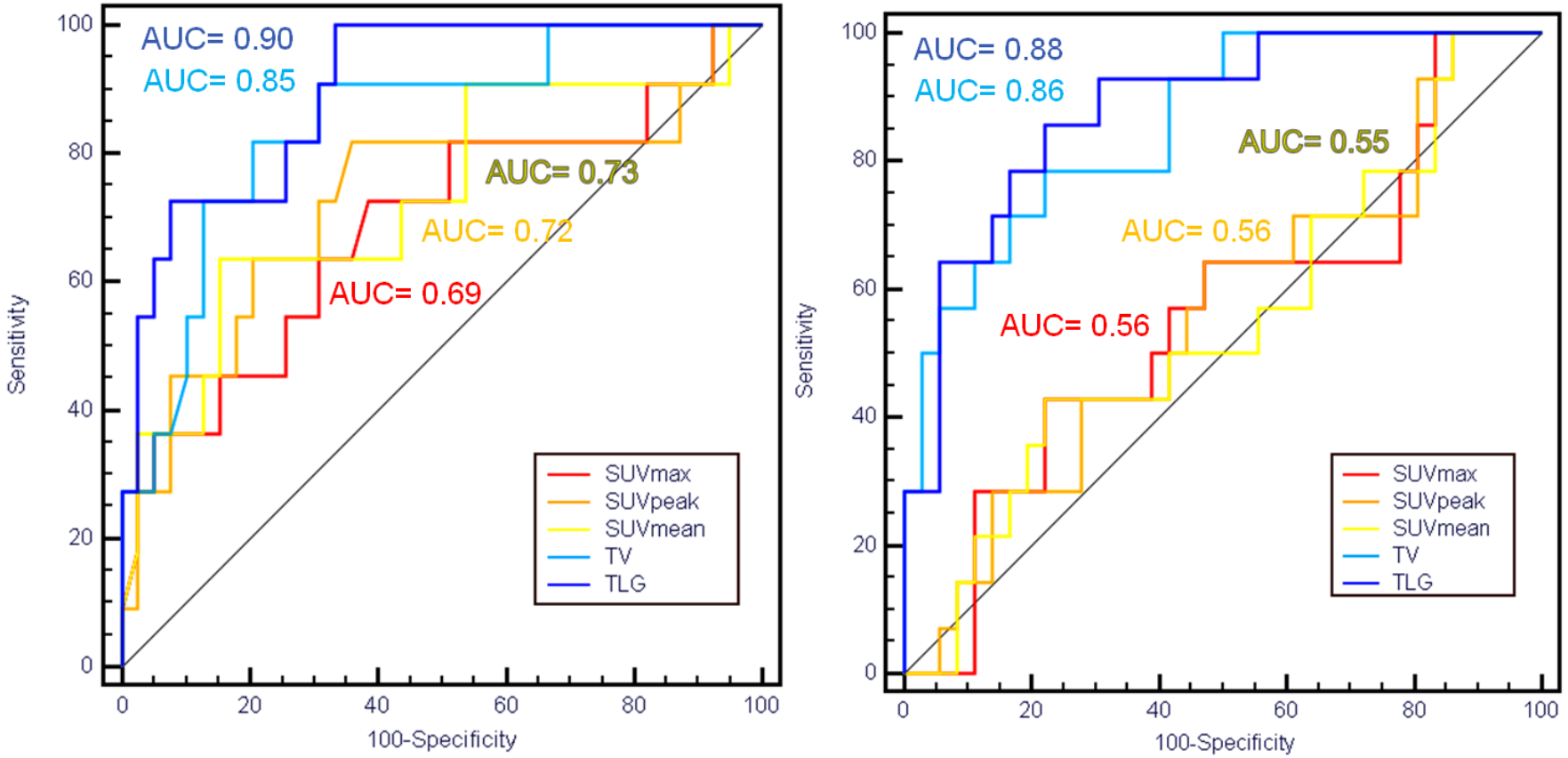



**Fig.2: identification of complete responders (left) and non responders (right) in locally advanced esophageal cancer using SUV (max, peak or mean), or tumor volume (TV) and associated total lesion glycolysis (TLG, volume multiplied by mean SUV) extracted from baseline $^{18}$F-FDG PET scans.**

b) Tumor targeting and dosimetry in radiotherapy

The use of CT images is considered the gold standard for the definition of tumor target volumes in radiotherapy planning (22). There is however a growing interest in including PET-based target delineation in order to reduce inter and intra observer delineation variability especially in cases where the tumor morphological contours are not clearly distinguishable (23). Also, the use of PET in radiotherapy is being considered to derive modified treatment plans using boosting or redistribution of the dose to specific parts of the tumor identified as being more metabolically active (using FDG), proliferative (using FLT) or hypoxic (using CU-ATSM, HX4 or FMISO), in order to reach higher tumor control probability (TCP) (13). However, this potentially requires accurate delineation of tumor volumes and sub-volumes in PET images. We therefore conducted different studies to demonstrate the dosimetry impact of such accurate delineation on standard IMRT (intensity-modulated radiotherapy) plans as well as the interest of dose redistribution or boosting, based on PET images with different tracers. Most of this work was conducted under my supervision by a PhD student (A. Le maitre), building on our previous work regarding advanced Monte Carlo simulations (24). We used such simulated datasets to compare delineation results by several approaches (FLAB and threshold-based methods) in terms of volume errors, sensitivity and positive predictive value. In addition, we investigated target coverage, dose homogeneity and D95 (target volume receiving 95% of prescribed dose), with a specific focus on cases with heterogeneous tracer distribution [1]. In addition, we carried out a study on simulated and clinical datasets regarding the impact of contrast and size of tumor functional sub-volumes on the dose redistribution and dose boosting scenarios, demonstrating the TCP improvement using boosting if contrast between the sub-volumes within the tumor is sufficient [2]. Finally, in collaboration with the MAASTRO clinic in the Netherlands, I also recently investigated the impact of tumor size and tracer uptake heterogeneity on the gross target volume delineation of non-small cell lung cancer validated by histopathology data on surgical specimens. The results highlighted a significant correlation between morphological volume and FDG uptake level of heterogeneity, and confirmed the need for advanced segmentation algorithms to obtain accurate PET heterogeneous delineation of the target volume [3].

**[1] A. Le Maître, D. Visvikis, C. Cheze-le Rest, O. Pradier, M. Hatt, Impact of the accuracy of automatic tumor functional volume delineation on radiotherapy treatment planning, Med Phys 2012, submitted.**

**[2] A. Le Maître, D. Visvikis, C. Cheze-le Rest, O. Pradier, M. Hatt, Dose prescription adapted to functional tumor radiotracer heterogeneities: the influence of contrast, PMB 2012, in revision.**

**[3] M. Hatt, et al, Impact of tumor size and tracer uptake heterogeneity in $^{18}$F-FDG PET and CT Non–Small Cell Lung Cancer tumor delineation, JNM 2011**

## C. Image fusion and paradoxical theory for therapy follow-up using sequential PET/CT images and multi-tracer PET images analysis

It has been proposed to consider the analysis of sequential PET/CT scans carried out before, during and after treatment in order to monitor response to therapy. However until now, most of the studies have only considered the evolution of simple measurements like $SUV_{max}$ (25). In addition, the use of multiple radiotracers to image different physiological processes (glucose consumption, cellular proliferation, hypoxia...) leads to as many images as tracers for a single patient (26). In both applications, the independent visual or semi-automatic analysis of each image might be insufficient as the correlation between images would not be fully exploited. In the work of a PhD student (S. David) under my supervision, we developed methods to automatically analyze and process multiple co-registered images, either sequential PET images for therapy follow-up, or the combination of different tracers, for instance to devise a biological target tumor volume in radiotherapy. The methods developed are based on multi observation fusion and classification within a Bayesian framework (27), in order to take into account all available information simultaneously. The approach demonstrated improved results with respect to independent segmentations on simulated data as well as clinical pre and post treatment PET images [1]. Another approach was also recently developed, based on paradoxical theory (Denzert-Smarandache fusion) (28) for the fusion of locally estimated parameters of interest [2]. Although these methods require rigid registration of multiple PET images and are therefore dependent on issues such as respiratory motion or anatomical changes during time, they could provide visual and quantitative estimation of tumor evolution during treatment as well as multi tracer analysis for radiotherapy target volume definition, as illustrated in figure 3 next page.

**[1] S. David, D. Visvikis, C. Roux, M. Hatt, Multi observation PET image analysis for patient follow-up quantitation and therapy assessment, PMB 2011 [selected as editor's choice]**

**[2] S. David, D. Visvikis, G. Quellec, P. Fernandez, M. Allard, C. Roux, M. Hatt, Image change detection using paradoxical theory for patient follow-up quantitation and therapy assessment, IEEE TMI 201, in revision.**

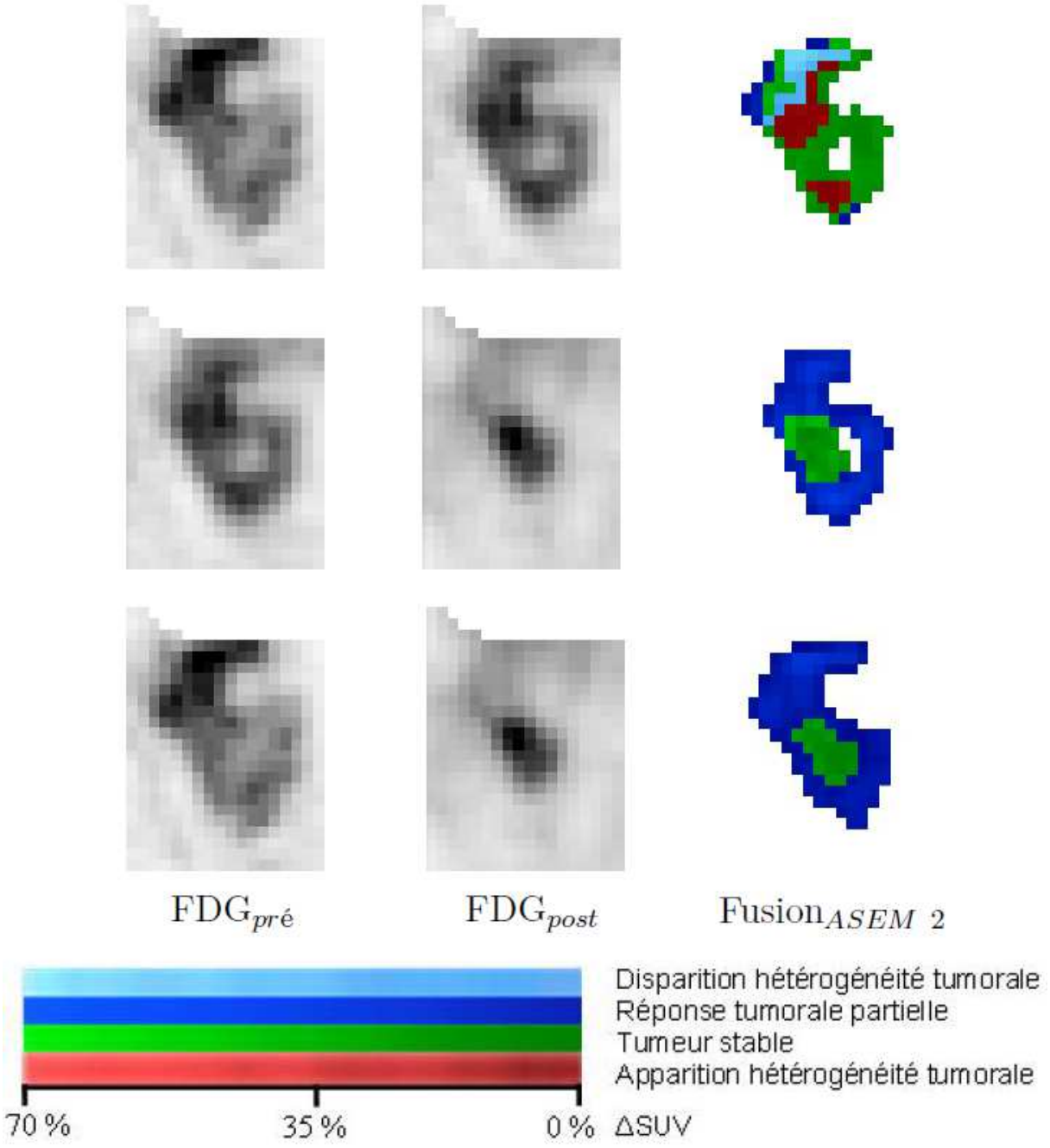



**Fig.3: pre et post treatment PET images of a necrotic rectal cancer tumor, with corresponding fusion showing responding (dark blue) or stable (green) voxels, as well as different sub-volume heterogeneities disappearing (light blue) or appearing (red).**

## D. PET image denoising and partial volume effects correction

I have been involved in the improvement of PET image wavelet-based denoising process by co-supervising the work of a PhD student (A. Le Pogam) as well as a master 2 trainee (H. Hanzouli). The major disadvantage of wavelets is the lack of direction information in the spatial-frequency transform, and therefore contours are not well preserved in the filtered image. On the contrary, curvelets explicitly model contours and better preserve them, whereas they are not as appropriate as wavelets to describe small point discontinuities (29). A method was then devised to combine wavelets and curvelets to reach the best compromise between denoising and preservation of the important information such as contours [1]. It should be emphasized that this method was also incorporated in the previously described deconvolution technique to improve the required denoising step. Using this improved deconvolution method, I recently demonstrated the lack of impact of PVE correction on prognostic or predictive clinical value of parameters derived from baseline PET scans in locally advanced esophageal cancer [2], although the impact of PVE correction on lesion detectability tasks or serial PET scans analysis remains to be evaluated.
I also contributed to another development of A. Le Pogam regarding the improvement of the partial volume effects correction method (MMA) developed in collaboration with N. Boussion and described previously (30). The first MMA method had two issues, first it was a global approach, meaning that the model linking functional and anatomical datasets was defined by one global parameter. The model was in addition applied to the entire image, independently of the correlation between datasets, which could introduce artefacts (such as bones from CT) in the corrected PET images. Second, it was a 2D method only. The method was therefore improved by designing a local and 3D model, and we demonstrated similar or significantly improved quantitative correction and qualitative visual aspects of the corrected images, with no artefacts in case of lack of correlation between the datasets [3]. This paper made the cover of the september issue of Medical Physics (see illustration in figure 4).

**[1] A. Le Pogam, H. Hanzouli, M. Hatt, et al, A combined 3-D wavelet and curvelet approach for edge preserving denoising in emission tomography, IEEE TMI 2012, submitted.**
**[2] M. Hatt, et al, Impact of partial volume effects correction on the predictive and prognostic value of baseline $^{18}$F-FDG PET images in esophageal cancer, JNM 2012, in press**
**[3] A. Le Pogam, M. Hatt, et al, Evaluation of a 3D local multi-resolution algorithm for the correction of partial volume effects in positron emission tomography, Med Phys 2011 [cover of the September issue]**

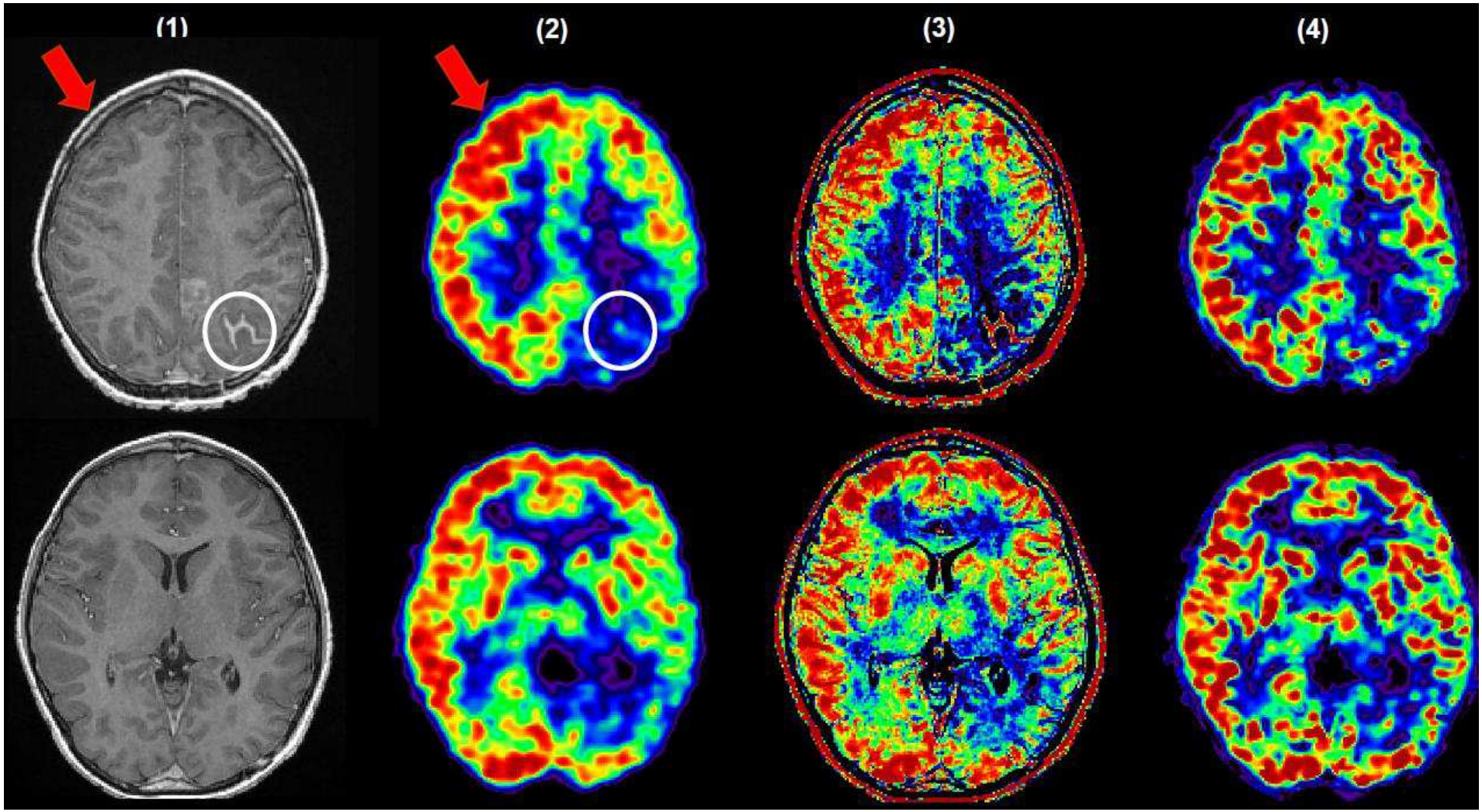



**Fig.4: PVE correction obtained with the original 2D MMA (3) versus the improved 3D local MMA (4) from the uncorrected PET (2) using details in the associated MRI (1). Notice the lack of artefacts in (4) with respect to the skull (red arrow) and the gadolinium injection (white cercle) that are introduced in the corrected image by the global approach (3).**

## VII - Future research project

Nowadays, the current trend in medical imaging is providing more and more images to characterize pathologies. Several imaging modalities have been used for decades (US, MRI, CT, PET), and new modalities are developed (optical imaging for instance). Existing modalities have also been expanded with new modes of acquisitions, providing complementary information, such as for instance different radiotracers in PET imaging beyond glucose metabolism, or various sequences with MRI. This trend has also been further emphasized since PET/CT multi modal devices have been introduced in the clinical setting for a decade now, and today with emerging PET/MRI systems, either sequential or simultaneous (31). The fourth dimension is also being more and more available, with significant advances in both hardware and software, allowing dynamic acquisitions to provide information on organs and tumors motion in both anatomical and functional modalities (32). It should also be emphasized that the dynamic nature of the acquisition can in addition be considered regarding the kinetics of tracers or contrast agents injected to the patients (33). In the near future, it may become more and more routine practice to carry out 5D acquisitions taking into account both temporal properties (physiological and morphological motion in addition to tracer kinetics). Thanks to less and less invasive procedures, another current trend in medical imaging consists of multiple acquisitions during the course of treatment, which may allow adaptive and potentially improved therapy in a variety of cancers (34-40).

Clinicians now have access to a large array of imaging modalities and devices providing complementary information and various views of the human body, on both morphological and physiological levels. In addition, these image datasets are almost systematically in three dimensions and full of details, therefore rather complex and time-consuming to fully analyze. The limitation of visual and manual analysis of one single 3D image dataset has already been underlined in numerous studies for specific tasks, such as for instance manual delineation of organs or tumors or detectability tasks. This led to restricting the majority of their clinical use to diagnosis or staging purposes in oncology, or sub-optimal treatment planning in radiotherapy. Such difficulties are exponentially increasing with the availability of additional 3D datasets of different imaging modalities, and/or additional dimensions to consider (motion, kinetics, and comparison of datasets during the course of treatment). A comprehensive, robust, reproducible and fast analysis of such large image datasets for a single patient or a patients' cohort would be impossible without semi-automated dedicated tools. The first contributions to the field in my research project will therefore be new and innovate approaches for the semi-automated processing and analysis of multi modal, multi resolution, multi dimensional images datasets.

Although primordial in today medicine, especially in oncology and radiotherapy, imaging is not the only source of information physicians base their decisions on. Similarly to the current increase of image-based data, another trend in medicine is the increase of the amount of data beyond imaging. Additional clinical information and data such as genetics or tumor biology are available and need to be taken into account. This means that this information may not only be correlated with imaging, but also combined with imaging for

increased predictive and prognostic power. A more personalized, preventive and predictive medicine for the future may benefit from decision aid systems combining the available data, both clinical and image-based (41-44). Additional contributions to the oncology and radiotherapy fields in my project will therefore be the development of a multi source fusion information framework to combine imaging and other clinical data into predictive and prognostic models for clinically reliable decision tools. This will require large datasets and the use of combined databases, which depends on local, regional, national and international research collaborations.

## 1. Multi observation framework for multi modal medical imaging

This part of my research project will be dedicated to the design and development of a robust and flexible mathematical and computational framework allowing to process or analyze complex, multi dimensional, multi resolution, multi modal co-registered images datasets. The goal is to design a framework that is flexible and robust enough to be able to carry out and implement the following fully or semi-automated complex image processing procedures on various image datasets such as for instance multi tracer PET images, PET/MRI or PET/CT datasets, including multiple scans during treatment and/or dynamic acquisitions: filtering and resolution recovery, segmentation and delineation, classification and pattern recognition.

I intend to develop such a framework by building on existing statistical hierarchical models such as Markov trees. I will focus on two major developments:

1. The multi observation framework

Markov trees can be used in order to include several images (or observations) within the same structure, instead of analyzing or processing each image separately (27). This requires prior co-registration, which could be obtained with high accuracy especially since multiple modalities are now often acquired almost (PET/CT) or completely simultaneously (integrated PET/MRI). Images with different resolutions can be included in the hierarchical model on appropriate scales, whereas images with similar resolution can also be included on the same scale, but with multi observation vectors associated to each node of the tree. Note that additional data could be taken into account in such a framework: wavelet decompositions of the images, annotations from physicians, various textural features images associated to each dataset, etc.

2. The modeling of correlations

One advantage of using such Markov tree models is that it should be possible to include in the model various correlations existing within or between the different images (see figure 5). Within each image, spatial correlations would be modeled by the intrinsic structure of the tree, and additional spatial correlation might be added. Similarly, correlations between images of different resolutions could be exploited by associating data of different resolution to different scales of the hierarchical model. Correlation between different modalities or different images at the same level of resolution could be estimated and used. This would be achieved thanks to the multi observation modeling, with a vector of several values

corresponding to the different images being associated with each node of the tree. The different types of noise and partial volume effects could be efficiently handled both within each image and in relationship with the other images included in the framework. This would be achieved via appropriate non stationary and correlated noise models (45). Such a framework would also be robust versus potential misregistration errors between datasets and missing data due to the modeling and exploitation of all these correlations.

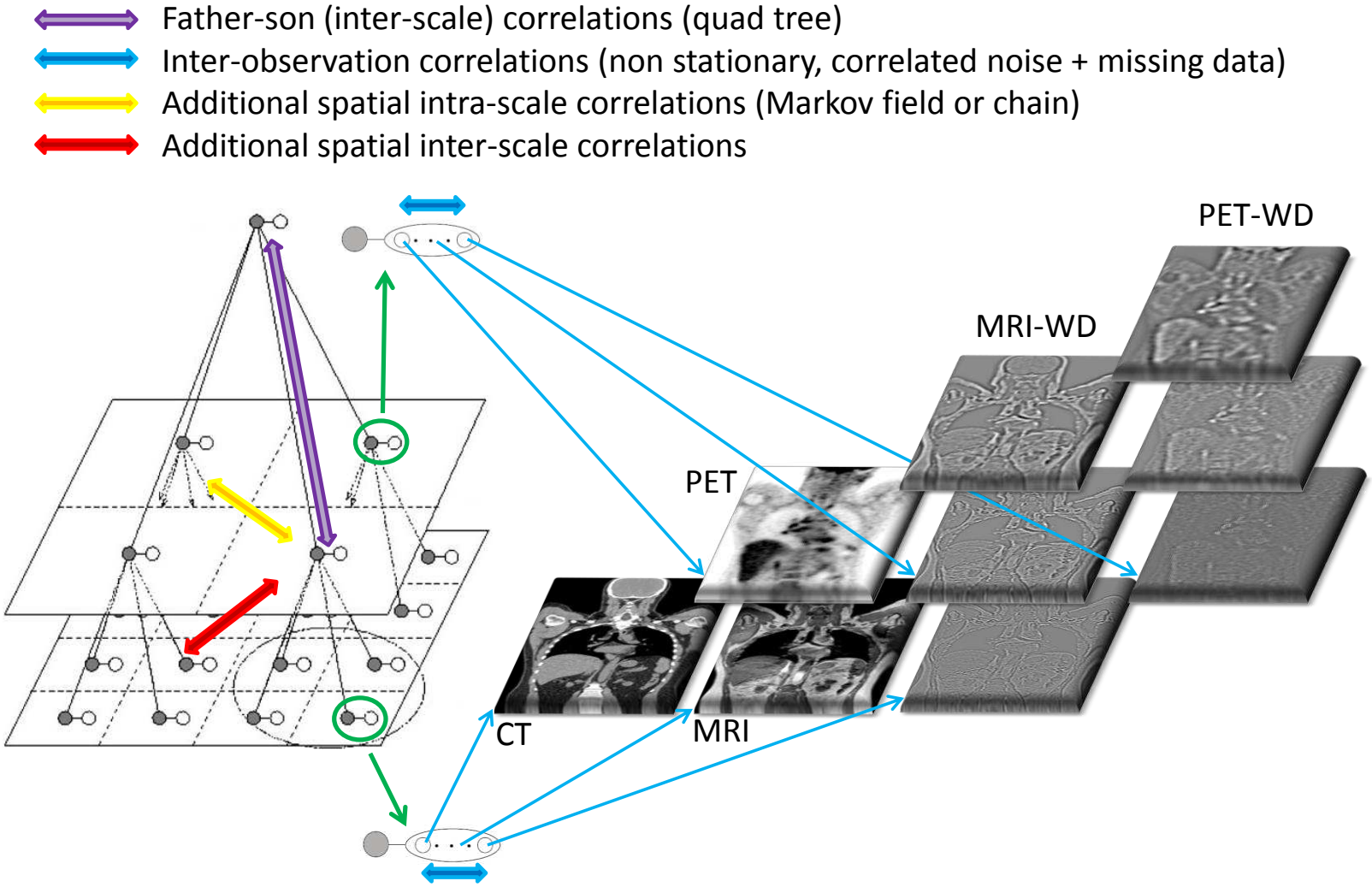



**Fig.5: illustration of the multi observation, multi scale framework**

The figure 5 on the previous page illustrates the potential of such structures for the proposed flexible framework. Note that this figure illustrates a Markov tree (a 3D structure) applied to 2D images with 16 (4x4) leaves and three scales (the root, a scale at 4 voxels and the last scale with 16), but it could in practice be extended to 3D datasets and obviously with a number of leaves adapted to the number of voxels in the attached images.

Grey circles represent the nodes of the tree (prior model) whereas white ones represent attached data (observed values in the images or other data such as wavelet decompositions, textural features, and so forth). As pointed out by the green arrow, this attached data may be multi observation, with various images or associated data included (see blue arrows linking elements of the observation vector to various images).

Large double arrows represent the various correlations that could be modeled and used for various applications. Blue ones denote correlations within the observations vector and could be implemented as various noise models. Basic modeling assumes independent observation and Gaussian noise, although more complex models could be used, such as non stationary, correlated, non Gaussian noise. It may also allow for missing data to be handled, since observation vectors may not contain the same data in each scale of the tree. Purple ones denote father-son statistical correlations linking elements of different scales in the quadtree that can add better handling of missing data and may relate information of images with different resolution (in the example above, a PET image is related to its associated CT and MRI datasets who have about four times more voxels). Yellow and red ones indicate additional spatial intra and inter scale correlations respectively that could be added in such a model to increase the robustness (as well as the complexity) of the model. As previously explained, data/images of different spatial resolutions may be associated with appropriate scales in the tree structure.

In this example, the leaves (the elements at the base of the quadtree) are associated with a vector of observation consisting of an MRI and a CT image (of approximately the same resolution). In the scale level above, a PET image (with about four times less voxels) is associated. WD denotes wavelet transforms. As these spatial-frequency transforms generates multi scale (from finer to coarser details as we go “up” in the tree structure) information, such additional data could also be attached as observations in the vector attached to the tree, or used as an guiding process, for instance within the context of couple or triplet Markov models that generalize standard Markov models allowing for more realistic modeling of real data (46).

Such models would also require dedicated developments of the associated parameters estimation procedures so the framework is automated enough to be used for applications such as image processing or analysis of large multi modal image datasets. The more complex and flexible the framework, the more parameters will need to be estimated in case of application to real data with unknown parameters. Robust algorithms such as Stochastic Expectation Maximization (SEM) or Iterative Conditional Estimation (ICE) will be adapted to the developed models (47).

As first applications of these developments, specific approaches will be investigated. They will be dedicated to multimodal PET/MRI and PET/CT, as well as dynamic imaging or sequential multimodal images during treatment. They will include automatic tumor localization/detection, improved denoising and partial volume effects correction, tumors and organs semi-automatic delineation, and static and dynamic parameters extraction to characterize pathologies. These approaches will be implemented within such a framework and are expected to benefit from its modeling versatility.

## 2. Multi source information fusion for predictive models and computer-aided decision in oncology

This part of my research project will be dedicated to the design and development of decision aid systems based on the exploitation of patients databases in various cancer models (for example esophageal, lung, or rectal) with known outcome (pathological response, disease-free and overall survival, etc.) in combination with clinical data and image-based parameters extracted thanks to the framework described above.

1. Multi modal image derived parameters obtained through developments carried out in the first part

As a first application of the multi observation framework described in part 1, I will implement automated multi modal characterization of tumors in oncology. Parameters such as anatomico-functional tumor volumes and associated measurements in various image modalities (SUVs of various PET tracers, heterogeneity of the tracer uptake or of the tissues in CT or MRI, diffusion, perfusion, dynamic information, etc.) could be extracted from large multi modal dataset in order to characterize fully the pathology in head and neck, esophageal, lung or rectal cancers. In addition, the temporal evolution of this full tumor characterization during treatment will also be of interest.

2. Clinical data including genetics and tumor biology

Fusion information (Denzert-Smarandache theory) (28) and classification approaches such as decision trees or support vector machines (49) will be investigated and compared on multi source data in order to derive predictive and prognostic models regarding each pathology for which patients databases are available. The goal will be to reach pertinent fusion of image-derived parameters and other additional semantic information such as clinical data (age, gender, stage…), tumor biology and genetics (from biopsies or histopathological examination, phenotypes, genotypes, etc.) as illustrated in figure 6. Such model learning requires large databases to avoid over fitting of the data, and multiple research collaborations will be needed with national and international clinical centers. I will exploit existing collaborations with research clinical teams in the Netherlands (Maastricht, Nijmegen), the United Kingdom (London, Surrey), the United States (MD Anderson), Germany (Freiburg, Munich) and France (Rennes, Nantes, Toulouse, Bordeaux, Brest) to help building such databases. Developed predictive and prognostic models will then be validated on prospective studies.

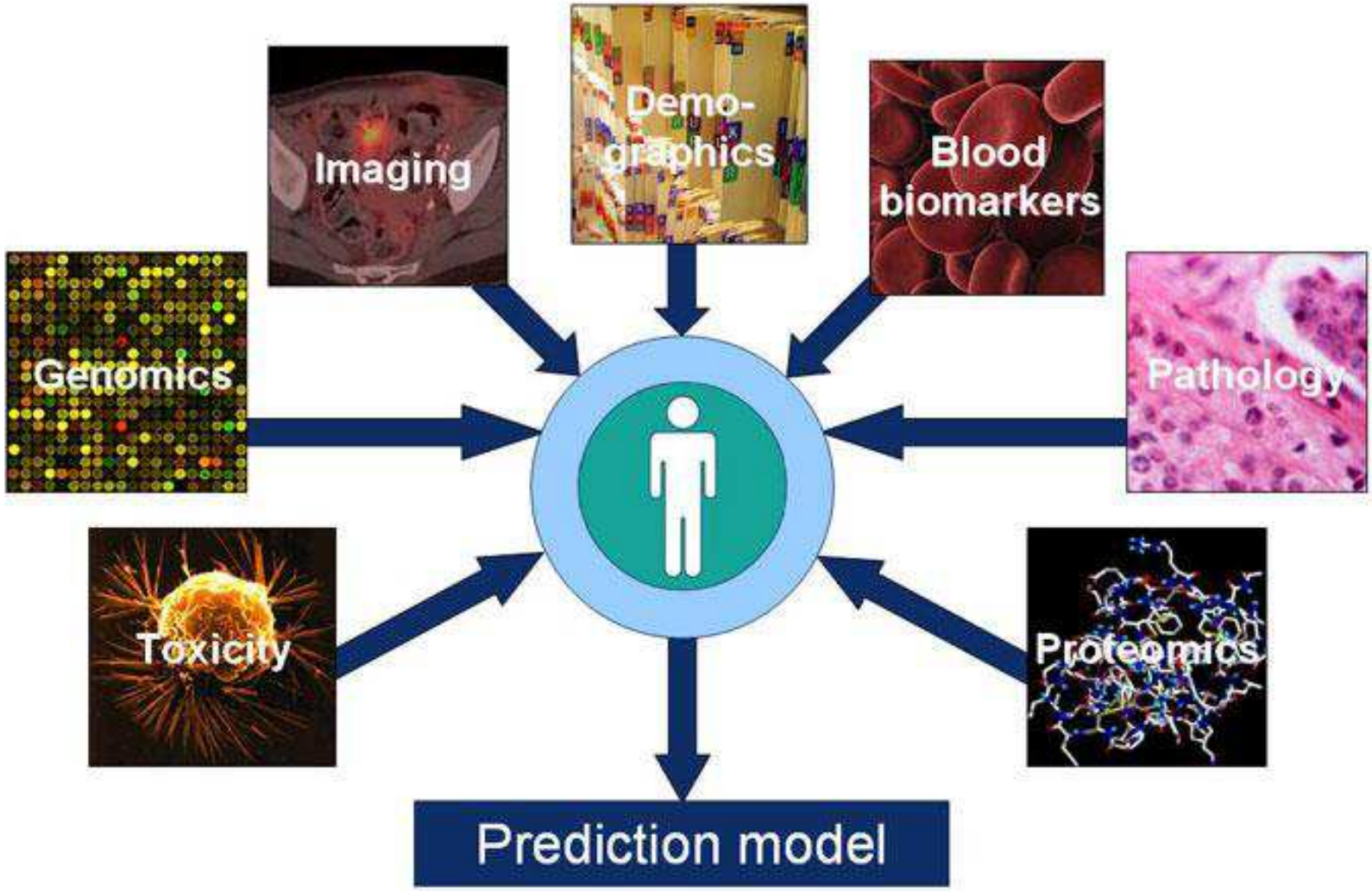



**Fig.6 : fusion of various data to build predictive models**

## VIII -Conclusion

I now have a significant experience in modeling for PET and PET/CT imaging applications in oncology and radiotherapy, with a special focus on semi-automated delineation and image processing algorithms. I have also significantly contributed to developments in partial volume effects correction, denoising, image fusion, radiotracer heterogeneity characterization and realistic Monte Carlo simulations and dosimetry studies in radiotherapy. I have further investigated the impact of such methodological developments in the clinical setting and associated applications. Most of these developments have involved co-supervision of several PhD students (in addition to several master students), two of them being supervised mostly by me and have now finished their PhD. My project for the next years involves the full supervision of two additional PhD students and co-supervision of at least two others.

## IX - References


**1.** Wahl RL, Jacene H, Kasamon Y, Lodge MA. From RECIST to PERCIST: Evolving Considerations for PET response criteria in solid tumors. J Nucl Med. 2009;50 Suppl 1:122S-150S.

**2.** Jarritt PH, Carson KJ, Hounsell AR, Visvikis D. The role of PET/CT scanning in radiotherapy planning. Br J Radiol. 2006;79 Spec No 1:S27-35.

**3.** Soret M, Bacharach SL, Buvat I. Partial-volume effect in PET tumor imaging. J Nucl Med. 2007;48:932-945.

**4.** Boellaard R. Standards for PET image acquisition and quantitative data analysis. J Nucl Med. 2009;50 Suppl 1:11S-20S.

**5.** Hatt M, Cheze-Le Rest C, Aboagye EO, et al. Reproducibility of 18F-FDG and 3'-deoxy-3'-18F-fluorothymidine PET tumor volume measurements. J Nucl Med. 2010;51:1368-1376.

**6.** Dewalle-Vignion A, Abiad AE, Betrouni N, Hossein-Foucher C, Huglo D, Vermandel M. Les méthodes de seuillage en TEP : un état de l'art Médecine Nucléaire. 2010;34:119-131.

**7.** Hatt M, Boussion N, Cheze-Le Rest C, Visvikis D, Pradier O. [Metabolically active volumes automatic delineation methodologies in PET imaging: Review and perspectives.]. Cancer Radiother. 2011.

**8.** Hatt M, Visvikis D. Defining radiotherapy target volumes using 18F-fluoro-deoxy-glucose positron emission tomography/computed tomography: still a Pandora's box?: in regard to Devic et al. (Int J Radiat Oncol Biol Phys 2010). Int J Radiat Oncol Biol Phys. 2010;78:1605.

**9.** Caillol H, Pieczynski W, Hillion A. Estimation of fuzzy Gaussian mixture and unsupervised statistical image segmentation. IEEE Trans Image Process. 1997;6:425-440.

**10.** Lucy LB. An iterative technique for the rectification of observed distributions. Astronomical Journal, Vol 79. 1974:745 (1974).

**11.** Richardson WH. Bayesian-Based Iterative Method of Image Restoration. Journal of the Optical Society of America, vol 62, issue 1. 1972:55.

**12.** Chang SG, Yu B, Vetterli M. Adaptive wavelet thresholding for image denoising and compression. IEEE Trans Image Process. 2000;9:1532-1546.

**13.** Supiot S, Lisbona A, Paris F, Azria D, Fenoglietto P. ["Dose-painting": myth or reality?]. Cancer Radiother. 2010;14:554-562.

**14.** Hofman MS, Hicks RJ. Restaging: should we percist without pattern recognition? J Nucl Med. 2010;51:1830-1832.

**15.** Buckler AJ, Boellaard R. Standardization of quantitative imaging: the time is right, and 18F-FDG PET/CT is a good place to start. J Nucl Med. 2011;52:171-172.

**16.** Nahmias C, Wahl LM. Reproducibility of standardized uptake value measurements determined by 18F-FDG PET in malignant tumors. J Nucl Med. 2008;49:1804-1808.

**17.** Velasquez LM, Boellaard R, Kollia G, et al. Repeatability of 18F-FDG PET in a multicenter phase I study of patients with advanced gastrointestinal malignancies. J Nucl Med. 2009;50:1646-1654.

**18.** Frings V, de Langen AJ, Smit EF, et al. Repeatability of metabolically active volume measurements with 18F-FDG and 18F-FLT PET in non-small cell lung cancer. J Nucl Med. 2010;51:1870-1877.

**19.** Lucignani G, Larson SM. Doctor, what does my future hold? The prognostic value of FDG-PET in solid tumours. Eur J Nucl Med Mol Imaging. 2010;37:1032-1038.

**20.** Basu S, Kwee TC, Gatenby R, Saboury B, Torigian DA, Alavi A. Evolving role of molecular imaging with PET in detecting and characterizing heterogeneity of cancer tissue at the primary and metastatic sites, a plausible explanation for failed attempts to cure malignant disorders. Eur J Nucl Med Mol Imaging. 2011.

**21.** Chua S, Dickson J, Groves AM. PET imaging for prediction of response to therapy and outcome in oesophageal carcinoma. Eur J Nucl Med Mol Imaging. 2011.

**22.** Chiti A, Kirienko M, Gregoire V. Clinical use of PET-CT data for radiotherapy planning: what are we looking for? Radiother Oncol. 2010;96:277-279.

**23.** Thorwarth D, Geets X, Paiusco M. Physical radiotherapy treatment planning based on functional PET/CT data. Radiother Oncol. 2010;96:317-324.

**24.** Le Maitre A, Segars W, Marache S, et al. Incorporating Patient-Specific Variability in the Simulation of Realistic Whole-Body 18F-FDG Distributions for Oncology Applications. Proceedings of the IEEE 2009;9:2026-2038.

**25.** Janssen MH, Ollers MC, van Stiphout RG, et al. PET-based Treatment Response Evaluation in Rectal Cancer: Prediction and Validation. Int J Radiat Oncol Biol Phys. 2011.

**26.** Janssen MH, Aerts HJ, Buijsen J, Lambin P, Lammering G, Ollers MC. Repeated positron emission tomography-computed tomography and perfusion-computed tomography imaging in rectal cancer: fluorodeoxyglucose uptake corresponds with tumor perfusion. Int J Radiat Oncol Biol Phys. 2011.

**27.** Pieczynski W. Modèles de Markov en traitement d'images. Traitement du Signal. 2003;20:255-278.

**28.** F. Smarandache JD, ed. Advances and Applications of DSmT for Information Fusion (Collected works): American Research Press; 2004-06.

**29.** Alzubi S, Islam N, Abbod M. Multiresolution analysis using wavelet, ridgelet, and curvelet transforms for medical image segmentation. Int J Biomed Imaging. 2011;2011:136034.

**30.** Boussion N, Hatt M, Lamare F, et al. A multiresolution image based approach for correction of partial volume effects in emission tomography. Phys Med Biol. 2006;51:1857-1876.

**31.** Sauter A, Kolb A, Soekler M, et al. Letter to the editor re: molecular imaging in oncology: the acceptance of PET/CT and the emergence of MR/PET imaging. Eur Radiol. 2011;21:1709-1712.

**32.** Aristophanous M, Berbeco RI, Killoran JH, et al. Clinical Utility of 4D FDG-PET/CT Scans in Radiation Treatment Planning. Int J Radiat Oncol Biol Phys. 2011.

**33.** Roe K, Aleksandersen TB, Kristian A, et al. Preclinical dynamic 18F-FDG PET - tumor characterization and radiotherapy response assessment by kinetic compartment analysis. Acta Oncol. 2010;49:914-921.

**34.** van Heijl M, Omloo JM, van Berge Henegouwen MI, et al. Fluorodeoxyglucose positron emission tomography for evaluating early response during neoadjuvant chemoradiotherapy in patients with potentially curable esophageal cancer. Ann Surg. 2011;253:56-63.

**35.** Aarntzen EH, Srinivas M, De Wilt JH, et al. Early identification of antigen-specific immune responses in vivo by [18F]-labeled 3'-fluoro-3'-deoxy-thymidine ([18F]FLT) PET imaging. Proc Natl Acad Sci U S A. 2011;108:18396-18399.

**36.** Vanderhoek M, Juckett MB, Perlman SB, Nickles RJ, Jeraj R. Early assessment of treatment response in patients with AML using [(18)F]FLT PET imaging. Leuk Res. 2011;35:310-316.

**37.** Rueger MA, Ameli M, Li H, et al. [18F]FLT PET for non-invasive monitoring of early response to gene therapy in experimental gliomas. Mol Imaging Biol. 2011;13:547-557.

**38.** Herrmann K, Buck AK, Schuster T, et al. A Pilot Study to Evaluate 3'-Deoxy-3'-18F-Fluorothymidine PET for Initial and Early Response Imaging in Mantle Cell Lymphoma. J Nucl Med. 2011;52:1898-1902.

**39.** Keam B, Im SA, Koh Y, et al. Early metabolic response using FDG PET/CT and molecular phenotypes of breast cancer treated with neoadjuvant chemotherapy. BMC Cancer. 2011;11:452.

**40.** Cortes Romera M, Gamez Cenzano C, Caresia Aroztegui AP, et al. Utility of the PET-CT in the evaluation of early response to treatment in the diffuse large B-cell lymphoma. Preliminary results. Rev Esp Med Nucl. 2011.

**41.** van Stiphout RG, Lammering G, Buijsen J, et al. Development and external validation of a predictive model for pathological complete response of rectal cancer patients including sequential PET-CT imaging. Radiother Oncol. 2011;98:126-133.

**42.** Valentini V, van Stiphout RG, Lammering G, et al. Nomograms for predicting local recurrence, distant metastases, and overall survival for patients with locally advanced rectal cancer on the basis of European randomized clinical trials. J Clin Oncol. 2011;29:3163-3172.

**43.** Starmans MH, Fung G, Steck H, Wouters BG, Lambin P. A simple but highly effective approach to evaluate the prognostic performance of gene expression signatures. PLoS One. 2011;6:e28320.

**44.** Egelmeer AG, Velazquez ER, de Jong JM, et al. Development and validation of a nomogram for prediction of survival and local control in laryngeal carcinoma patients treated with radiotherapy alone: a cohort study based on 994 patients. Radiother Oncol. 2011;100:108-115.

**45.** Benboudjema D, Pieczynski W. Unsupervised statistical segmentation of nonstationary images using triplet Markov fields. IEEE Trans Pattern Anal Mach Intell. 2007;29:1367-1378.

**46.** Pieczynski W. Multi sensor triplet Markov chains and theory of evidence. International Journal of Approximate Reasoning. 2007;45:1-16.

**47.** Pieczynski W, Bouvrais J, Michel C. Estimation of generalized mixture in the case of correlated sensors. IEEE Trans Image Process. 2000;9:308-312.

**48.** Quellec G, Lamard M, Bekri L, Cazuguel G, Cochener B, Roux C. Multimedia medical case retrieval using decision trees. Conf Proc IEEE Eng Med Biol Soc. 2007;2007:4536-4539.

**49.** El Naqa I, Bradley JD, Lindsay PE, Hope AJ, Deasy JO. Predicting radiotherapy outcomes using statistical learning techniques. Phys Med Biol. 2009;54:S9-S30.

# X - Appendix (full list of communications, copies of published papers)

**Invited talks (international conferences)**

1. PET image derived indices for therapy monitoring applications: what may the future hold?, Society of nuclear medicine annual meeting, continuous education session *"PET image derived indices for therapy response and outcome prediction studies"*, San Antonio, USA, june 7, 2011.

2. Validation of tumor delineation algorithms dedicated to PET imaging, Society of nuclear medicine annual meeting, *continuous education session "PET* tumor *functional volume quantification"*, Salt Lake City, USA, june 9, 2010.

3. Segmentation of functional volumes in PET/CT, Quantitative imaging and dosimetry in Nuclear Medicine, Berder island, october 5, 2006.

**Invited talks (international institutions)**

1. PET images processing and analysis for oncology and radiotherapy applications: pitfalls, methods and results, seminar in Maastro clinic, imaging and radiotherapy research department*"*, Maastricht, the Netherlands, november 22, 2011.

2. Accuracy and robustness study of an automatic tumor delineation method in positron emission tomography for oncology applications, Hammersmith Hospital Imperial College, London, November 13, 2009.

3. Définition du volume cible métabolique en TEP/TDM: aspects méthodologiques, CHU de Liège, 23 juin 2009.

**Communications in conferences [posters or oral communications]**

**M. Hatt**, C. Cheze Le Rest, D. Visvikis, O. Pradier, La définition précise des volumes métaboliques sur TEP au $^{18}$F-FDG avant traitement permet la prédiction de la réponse à la chimio radiothérapie dans les cancers de l'œsophage, 22$^{ème}$ congrès de la SFRO, 2011 [oral, session "meilleurs résumés"].

**M. Hatt**, C. Cheze Le Rest, O. Pradier, D. Visvikis, Impact of Tumor Size and $^{18}$F-FDG Tracer Uptake Heterogeneity in Non–Small Cell Lung Cancer Tumor Automatic Delineation On PET and CT Images for Gross Tumor Volumes Determination, 53$^{rd}$ Joint AAPM/COMP Meeting, 2011 [oral & poster, award *best-in-physics*].

**M. Hatt**, D. Visvikis, O. Pradier, C. Cheze Le Rest, Tumor Metabolic Dimension Measurements On $^{18}$F-FDG PET Pre Treatment Image Predicts Response to Exclusive Concomitant Radiochemotherapy in Locally Advanced Esophageal Cancer, 53$^{rd}$ Joint AAPM/COMP Meeting, 2011 [oral].

**M. Hatt**, A. Le Pogam, D. Visvikis, O. Pradier, C. Cheze Le Rest, Predictive and Prognostic Clinical Values of $^{18}$F-FDG PET Based Measurements in Locally Advanced Esophageal Cancer Are Not Improved by Partial Volume Effects Correction, 53$^{rd}$ Joint AAPM/COMP Meeting, 2011 [Poster].

S. David, **M. Hatt**, D. Visvikis, Multi Observation PET Image Fusion for Patient Follow-Up Quantitation in Oncology, 53$^{rd}$ Joint AAPM/COMP Meeting, 2011 [Poster].

A. Lemaitre, D. Wallach, **M. Hatt**, S. Edel, N. Boussion, O. Pradier, D. Visvikis, Impact of 4D PET and Motion Correction in the Delineation of Gross Tumor Volume for Radiotherapy Treatment Planning in Lung Cancer, 53rd Joint AAPM/COMP Meeting, 2011 [Poster].

**M. Hatt**, D. Visvikis, O. Pradier, C. Cheze Le Rest, Total lesion glycolysis measured on $^{18}$F-FDG PET baseline scan predicts radiochemotherapy response in locally advanced esophageal cancer, Society of nuclear medicine annual meeting, 2011 [Poster].

T. Wentz, H. Fayad, J-F. Clément, J. Savean, **M. Hatt**, D. Visvikis, Extraction and evaluation of anatomical patient surface and associated respiratory motion with a Time-of-Flight (ToF) camera, Society of nuclear medicine annual meeting, 2011 [Oral].

F. Tixier, **M. Hatt**, C. Cheze Le Rest, D. Visvikis, Reproducibility of textural feature measurements extracted from $^{18}$F-FDG PET images for tumor heterogeneity characterization, Society of nuclear medicine annual meeting, 2011 [Poster].

F. Tixier, **M. Hatt**, C. Cheze Le Rest, D. Visvikis, Quantification of the intra-tumor heterogeneity on baseline $^{18}$F-FDG PET images characterized by textural features, Society of nuclear medicine annual meeting, 2011 [Poster].

F. Tixier, **M. Hatt**, D. Visvikis, C. Cheze Le Rest, Intra-tumor heterogeneity on baseline $^{18}$F-FDG PET images characterized by textural features predicts response to concomitant radio-chemotherapy in esophageal cancer, Society of nuclear medicine annual meeting, 2011 [Poster].

N. Withofs, C. Bernard, C. van der Rest, P. Martinive, **M. Hatt**, D. Visvikis, P. Coucke, R. Hustinx, FDG PET/CT for radiotherapy treatment planning. Comparison of functional volume delineation algorithms, Society of nuclear medicine annual meeting, 2011 [Poster].

A. Le Maitre, **M. Hatt**, C. Cheze Le Rest, O. Pradier, D. Visvikis, Impact of functional contrast on non-uniform radiotherapy dose prescriptions, 11th Biennial ESTRO on Physics & Radiation Technology For Clinical Radiotherapy, 2011. [Poster]

F. Tixier, **M. Hatt**, O. Pradier, L. Corcos, D. Visvikis, C. Cheze Le Rest, Intra-tumor heterogeneity on baseline $^{18}$F-FDG PET images characterized by textural features predicts response to concomitant radio-chemotherapy in esophageal cancer, 49ème Colloque de Médecine Nucléaire de Langue Française (SFMN), 2011. [Oral]

L. Ghouti, **M. Hatt**, F. Courbon, J. Selves, M. Poirot, R. Guimbaud, B. Pradère, Optimisation des méthodes d'évaluation des volumes tumoraux en TEP $^{18}$FDG et IRM. Corrélation à l'histo-imagerie quantitative. Application aux cancers du rectum post-radiochimiothérapie, Journées francophones *d'hépato-gastroentérologie et d'oncologie digestive,* 2011. [Poster]

**M. Hatt**, D. Visvikis, M. N. Albarghach, O. Pradier, C. Cheze Le Rest, Valeur prédictive et pronostique du volume fonctionnel des lésions en $^{18}$F-FDG TEP dans le cancer de l'œsophage, 5e Journées du Cancéropôle Grand Ouest, 2010. [Poster]

A. Le Maitre, **M. Hatt**, C. Cheze Le Rest, O. Pradier, D. Visvikis, Adapting Dose Prescription to Tumor Heterogeneities: Influence of the Functional Contrast, IEEE Nuclear Science Symposium and Medical Imaging Conference records 2010. [Poster]

S. David, **M. Hatt**, N. Boussion, P. Fernandez, M. Allard, O. Barrett, D. Visvikis, A Multi-Observation Fusion Approach for Patient Follow-up Using PET/CT, IEEE Nuclear Science Symposium and Medical Imaging Conference records 2010. [Poster]

**M. Hatt**, M. N. Albarghach, D. Visvikis, C. Cheze Le Rest, Prognostic value of Total Glycolytic Volume in esophagus cancer: impact of automatic tumor volume delineation on $^{18}$F-FDG PET images, Imaging for treatment assessment in radiation therapy, 2010. [Oral]

A. Le Maître, **M. Hatt**, M. N. Albarghach, O. Pradier, C. Cheze Le Rest, M. Bal, D. Visvikis, Dosimetry impact of accurate PET segmentation for radiotherapy treatment planning, Imaging for treatment assessment in radiation therapy, 2010. [Oral]

**M. Hatt**, M. N. Albarghach, D. Visvikis, C. Cheze Le Rest, Automatic $^{18}$F-FDG PET tumor volume delineation and prognostic value of Total Glycolytic Volume in esophagus cancer, Journal of Nuclear Medicine 2010:51(S2):110. [Oral]

**M. Hatt**, M. N. Albarghach, C. Cheze Le Rest, D. Visvikis, Valeur prognostique du volume tumoral actif en $^{18}$F-FDG pour le cancer de l'œsophage et influence de la méthodologie de contourage de la tumeur, 48ème colloque de médecine nucléaire de langue française, 2010. [Poster]

**M. Hatt**, C. Cheze Le Rest, D. Visvikis, Robustness and reproducibility of PET functional volumes automated delineation: comparison of various approaches, Molecular Imaging in Radiation Oncology, 2010. [Poster]

A. Le Maître, **M. Hatt**, M. N. Albarghach, C. Cheze Le Rest, O. Pradier, D. Visvikis, Impact of the accuracy of tumor functional volume delineation on radiotherapy treatment planning, Molecular Imaging in Radiation Oncology, 2010. [Poster]

**M. Hatt**, C. Cheze Le Rest, O. Pradier, D. Visvikis, Is accurate and reproducible biological target volumes delineation in PET images for radiotherapy planning feasible?, Institute of Physics and Engineering in Medicine – uses of PET in radiotherapy, 2010. [Oral]

M. Lei, J. Crawshaw, J. Scuffham, D. Rickard, **M. Hatt**, J. Hall, J. Sellinger, T. Jordan, S. Whitaker, D. Visvikis, A. Nisbet, T. Guerrero Urbano, Biological gross tumor volume definition in head and neck squamous cell carcinoma (HNSCC) radiotherapy planning: Comparison of automated segmentation tools, Institute of Physics and Engineering in Medicine – uses of PET in radiotherapy, 2010. [Oral]

**M. Hatt**, C. Collet, F. Salzenstein, C. Roux, D. Visvikis, From nebulae segmentation in astronomical imaging to tumor delineation in $^{18}$F-FDG PET imaging: how can one serve the other?, AstroMed09: The Inaugural Sydney International Workshop on Synergies in Astronomy and Medicine, 2009. [Oral]

S. David, **M. Hatt**, D. Visvikis, Multi-tracer image fusion for patients using multi-observation statistical image fusion framework: a feasibility study, AstroMed09: The Inaugural Sydney International Workshop on Synergies in Astronomy and Medicine, 2009. [Oral]

S. David, **M. Hatt**, N. Boussion, P. Fernandez, M. Allard, O. Barrett, D. Visvikis, Multi-Tracer PET Image Fusion Using Fuzzy Logic: a Feasibility Study, IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS-MIC), 2009. [Poster]

**M. Hatt**, D. Visvikis, C. Cheze Le Rest, O. Pradier, Définition automatique des volumes biologiques cibles pour les applications de radiothérapie, Cancer / Radiothérapie 2009;13(6-7):640-641. [Oral]

N. Albarghach, **M. Hatt**, D. Cornec, S. Querellou, C. Berthou, Y. Renaudineau, D. Visvikis, O. Pradier, C. Cheze Le Rest, Intérêt de la TEP au FDG dans les lymphomes folliculaires, Congrès National de la Société Française de Radiothérapie Oncologique (SFRO), 2009. [Poster]

D. Visvikis, **M. Hatt**, O. Pradier, N. Albarjhach, C. Cheze Le Rest, Use of an automatic tumor delineation algorithm for therapy response assessment with PET, Annual Congress of the European Association of Nuclear Medicine, 2009. [Oral]

**M. Hatt**, C. Cheze Le Rest, D. Visvikis, Accuracy and robustness study of a new automatic tumor delineation method in positron emission tomography for patients' follow up and therapy response assessment, New trends in molecular imaging and nuclear medicine – young investigators award, 2009. [Oral]

A. Le Pogam, **M. Hatt**, N. Boussion, L. Livieratos, A.M. Alessio, C. Cheze Le Rest and D. Visvikis, Comparison of voxel-wise partial volume correction approaches for emission tomography, New trends in molecular imaging and nuclear medicine – young investigators award, 2009. [Poster]

**M. Hatt**, A. Dekker, D. De Ruysscher, M. Oellers, P. Lambin, C. Roux, C. Cheze le Rest, Une nouvelle méthode pour la détermination automatique des volumes fonctionnels en oncologie : premiers résultats cliniques, 47ème colloque de médecine nucléaire de langue française, 2009. [Poster]

**M. Hatt**, A. Turzo, P. Bailly, I. Murray, C. Roux, D. Visvikis, Automatic delineation of functional volumes in PET: a robustness study, Journal of Nuclear Medicine 2009:50(S2):1429. [Poster]

**M. Hatt**, C. Cheze Le Rest, O. Pradier, D. Visvikis, Automatic PET tumor delineation for patient's follow-up and therapy assessment, Journal of Nuclear Medicine 2009:50(S2):182. [Oral]

A. Le Pogam, P. Descourt, **M. Hatt**, N. Boussion, D. Visvikis, A combined 3-D wavelet and curvelet approach for edge preserving denoising in emission tomography, Journal of Nuclear Medicine 2009;50(S2):533. [Oral]

A. Le Maitre, W.P. Segars, A. Reilhac, **M. Hatt**, S. Tomei, C. Lartizien, D. Visvikis, Incorporating patient specific variability in a database of realistic simulated whole body $^{18}$F-FDG distributions for oncology applications, Journal of Nuclear Medicine 2009;50(S2):1488. [Poster]

**M. Hatt**, A. Dekker, D. De Ruysscher, M. Oellers, P. Lambin, C. Roux, D. Visvikis, Olivier Pradier, Accurate and robust functional volume definition in PET for radiotherapy treatment planning, DEGRO 2009, 2009. [Oral]

**M. Hatt**, C. Cheze Le Rest, A. Dekker, D. De Ruysscher, M. Oellers, P. Lambin, C. Roux, D. Visvikis, Une nouvelle méthode automatique de détermination des volumes fonctionnels pour l'oncologie, Journées de Recherche en Imagerie et Technologies de la santé (RITS), 2009. [Oral]

A. Le Pogam, N. Boussion, **M. Hatt**, C. Prunier-Aesch, D. Guilloteau, J-L. Baulieu, D. Visvikis, Transformées en Ridgelet/Curvelet discrètes 3-D pour le débruitage en Tomographie d'Emission, Journées de Recherche en Imagerie et Technologies de la santé (RITS), 2009. [Oral]

**M. Hatt**, P. Bailly, A Turzo, C. Roux, D. Visvikis, PET functional volume segmentation: a robustness study, IEEE Nuclear Science Symposium and Medical Imaging Conference records 2008:4335-4339. [Poster]

**M. Hatt**, A. Dekker, D. De Ruysscher, M. Oellers, P. Lambin, C. Roux, D. Visvikis, Accurate functional volume definition in PET for radiotherapy treatment planning, IEEE Nuclear Science Symposium and Medical Imaging Conference records 2008:5567-5571. [Oral]

A. Le Pogam, **M. Hatt**, N. Boussion, F. Turkheimer, C. Prunier-Aesch, D. Guilloteau, J.-L. Baulieu, D. Visvikis, A wavelet-based hidden Markov model and multi-resolution approach for conditional

partial volume correction in emission tomography, IEEE Nuclear Science Symposium and Medical Imaging Conference records 2008:1-5. [Poster]

A. Le Pogam, **M. Hatt**, N. Boussion, D. Guilloteau, J-L. Beaulieu, C. Prunier, F. Turkheimer, D. Visvikis, A 3D multi resolution local analysis approach for correction of partial volume effects in emission tomography, IEEE Nuclear Science Symposium and Medical Imaging Conference records 2008:5300-5303. [Poster]

A. Le Pogam, N. Boussion, **M. Hatt**, C. Prunier-Aesch, D. Guilloteau, J.-L. Baulieu, D. Visvikis, 3D discrete ridgelet transform for emission tomography denoising, IEEE Nuclear Science Symposium and Medical Imaging Conference records 2008:5557-5561. [Oral]

**M. Hatt**, D. Visvikis, N. Boussion, O. Pradier, Segmentation automatique d'images fonctionnelles TEP pour l'aide à la planification de traitement en radiothérapie, 47e journées scientifiques de la Société Française de Physique Médicale (SFPM), 2008. [Poster]

**M. Hatt**, N. Boussion, C. Roux, O. Pradier, C. Cheze Le Rest, D. Visvikis, An automatic segmentation algorithm for accurate functional volume determination in radiotherapy treatment planning, Annual Congress of the European Association of Nuclear Medicine, 2008. [Oral]

**M. Hatt**, N. Boussion, O. Pradier, C. Roux, D. Visvikis, Automatic Segmentation of Functional Images for Radiotherapy Treatment Planning, International Journal of Radiation Oncology Biology Physics 2008;72(1):S682. [Poster]

**M. Hatt**, C. Cheze Le Rest, C. Roux, D. Visvikis, Automatic volume delineation of heterogeneous tumors in PET: comparison of various methodologies, Journal of Nuclear Medicine 2008;49(S1):381. [Poster]

N. Boussion, C. Cheze-Le Rest, **M. Hatt**, D. Visvikis, Evaluation of resolution and quantitation preserving wavelet-based denoising in wholebody PET, Journal of Nuclear Medicine 2008;49(S1):395. [Poster]

N. Boussion, **M. Hatt**, D. Visvikis, Partial volume correction in PET based on functional volumes, Journal of Nuclear Medicine 2008;49(S1):388. [Poster]

A. Le Pogam, **M. Hatt**, N. Boussion, D. Guilloteau, J-L. Beaulieu, C. Prunier, F. Turkheimer, D. Visvikis, Conditional voxel-wise partial volume correction for emission tomography: combining a wavelet-based hidden Markov model with a mutual multi-resolution analysis approach, Journal of Nuclear Medicine 2008;49(S1):62. [Oral]

A. Le Pogam, **M. Hatt**, N. Boussion, D. Guilloteau, J-L. Beaulieu, C. Prunier, F. Turkheimer, D. Visvikis, Conditional partial volume correction for emission tomography: a wavelet-based hidden Markov model and multi-resolution approach, 5th IEEE International Symposium on Biomedical Imaging conference records 2008:1319-1322. [Oral]

N. Boussion, **M. Hatt**, D. Wallach, O. Pradier, D. Visvikis, Quantification en TEP oncologique : les dernières avancées en traitement d'images et leur impact potentiel sur la planification de traitement en radiothérapie, 47e journées scientifiques de la société française de physique médicale, Marseille France, juin 2008. [Oral]

**M. Hatt**, C. Roux, D. Visvikis, A Segmentation Algorithm for Heterogeneous Tumor Delineation in PET, IEEE Nuclear Science Symposium and Medical Imaging Conference records 2007;5:3939-3945. [Poster]

N. Boussion, **M. Hatt**, A. Reilhac, D. Visvikis, Fully automated partial volume attenuation in PET: a wavelet approach without anatomical information, IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS-MIC) conference records 2007;4:2812-2816. [Poster]

**M. Hatt**, C. Roux, D. Visvikis, Evaluation de méthodes de segmentation bayésiennes pour l'imagerie TEP en oncologie, *GRETSI, Groupe d'Etudes du Traitement du Signal et des Images*, 2007. [Poster]

**M. Hatt**, N. Boussion, C. Roux, D. Visvikis, Evaluation de méthodes de segmentation statistiques pour l'imagerie TEP en oncologie, Journée de Recherche en Imagerie Médicale (JRIM), 2007. [Poster]

N. Boussion, **M. Hatt**, F. Lamare, D. Visvikis, Correction des effets de volume partiel en TEP sans a priori anatomique : déconvolution et filtrage à base d'ondelettes, Journée de Recherche en Imagerie Médicale (JRIM), 2007. [Oral]

**M. Hatt**, C. Roux, D. Visvikis, 3D Fuzzy Adaptive Unsupervised Bayesian Segmentation for Volume Determination in PET, 4th IEEE International Symposium on Biomedical Imaging: From Nano to Macro conference records 2007:328-331. [Poster]

**M. Hatt**, C. Cheze-Le Rest, A. Turzo, F. Lamare, K. Carson, P. Jarritt, D. Visvikis, An automatic segmentation algorithm for functional volume and activity concentration determination in PET, Annual Congress of the European Association of Nuclear Medicine, 2006. [Oral]

**M. Hatt**, N. Boussion, F. Lamare, C. Collet, F. Salzenstein, C. Roux, Y. Bizais, C. Cheze-Le Rest, D. Visvikis, Fuzzy versus Hard Hidden Markov Chains Segmentation for Volume Determination and Quantitation in Noisy PET Images, 3rd IEEE International Symposium on Biomedical Imaging: From Nano to Macro conference records 2006:1376-1379. [Poster]

**M. Hatt** t, N. Boussion, K. Karson, P. Jarritt, F. Lamare, Y. Bizais, C. Cheze-Le Rest, D. Visvikis, Comparison of different methodologies for lesion volume determination in PET, Nuclear Medicine Communications 2006 27(3):283-284. [Oral]

A. Oberto, M. Wenger, J.-P. Lejal, S. Jaehn, B. Baranne, **M. Hatt**, O. Dellicour, J. Deprez. SIMBAD 4: a new Release with new Possibilities, Astronomical Data Analysis Software and Systems XV ASP Conference Series, Astronomical Society of the Pacific 2006;351:703-706. [Poster]

N. Boussion, **M. Hatt**, F. Lamare, Y. Bizais, A. Turzo, C. Cheze-Le Rest, D. Visvikis, Generating resolution-enhanced images for correction of partial volume effects in emission tomography: a multiresolution approach, IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS-MIC) Conference Records 2005;4:2423-2427. [Poster]

N. Boussion, **M. Hatt**, F. Lamare, A. Turzo, Y. Bizais, C. Cheze-Le Rest, D. Visvikis, A multiresolution approach in partial volume correction for emission tomography imaging, European Journal of Nuclear Medicine and Molecular Imaging 2005;32(S1):S51. [Oral]

Phys. Med. Biol. **52** (2007) 3467–3491 doi:10.1088/0031-9155/52/12/010

# Fuzzy hidden Markov chains segmentation for volume determination and quantitation in PET

**M Hatt[1], F Lamare[1], N Boussion[1], A Turzo[1,2], C Collet[3], F Salzenstein[4], C Roux[1,5], P Jarritt[6], K Carson[6], C Cheze-Le Rest[1,2] and D Visvikis[1]**

[1] INSERM U650, Laboratoire du Traitement de l'Information Médicale (LaTIM), CHU Morvan, Bat 2bis (I3S), 5 avenue Foch, Brest, 29609, France
[2] Academic Department of Nuclear Medicine, CHU Morvan, Brest, F-29609, France
[3] Ecole Nationale Supérieure de Physique de Strasbourg (ENSPS), ULP, Strasbourg, F-67000, France
[4] Institut d'Électronique du Solide et des Systèmes (InESS), ULP, Strasbourg, F-67000, France
[5] ENST Bretagne, GET-ENST, Brest, F-29200, France
[6] Medical Physics Agency, Royal Victoria Hospital, Belfast, UK

Received 4 August 2006, in final form 4 April 2007
Published 18 May 2007
Online at stacks.iop.org/PMB/52/3467

**Abstract**

Accurate volume of interest (VOI) estimation in PET is crucial in different oncology applications such as response to therapy evaluation and radiotherapy treatment planning. The objective of our study was to evaluate the performance of the proposed algorithm for automatic lesion volume delineation; namely the fuzzy hidden Markov chains (FHMC), with that of current state of the art in clinical practice threshold based techniques. As the classical hidden Markov chain (HMC) algorithm, FHMC takes into account noise, voxel intensity and spatial correlation, in order to classify a voxel as background or functional VOI. However the novelty of the fuzzy model consists of the inclusion of an estimation of imprecision, which should subsequently lead to a better modelling of the 'fuzzy' nature of the object of interest boundaries in emission tomography data. The performance of the algorithms has been assessed on both simulated and acquired datasets of the IEC phantom, covering a large range of spherical lesion sizes (from 10 to 37 mm), contrast ratios (4:1 and 8:1) and image noise levels. Both lesion activity recovery and VOI determination tasks were assessed in reconstructed images using two different voxel sizes (8 $mm^3$ and 64 $mm^3$). In order to account for both the functional volume location and its size, the concept of % classification errors was introduced in the evaluation of volume segmentation using the simulated datasets. Results reveal that FHMC performs substantially better than the threshold based methodology for functional volume determination or activity concentration recovery considering a contrast ratio of 4:1 and lesion sizes of $<28$ mm. Furthermore differences between classification and volume estimation errors evaluated were smaller for the segmented volumes provided by the FHMC algorithm. Finally, the performance of the automatic algorithms was less susceptible to image

noise levels in comparison to the threshold based techniques. The analysis of both simulated and acquired datasets led to similar results and conclusions as far as the performance of segmentation algorithms under evaluation is concerned.

## 1. Introduction

Positron emission tomography (PET) has been long established as a powerful tool in oncology, particularly in the area of diagnosis. However, alternative applications such as the use of PET in radiotherapy planning (Jarritt *et al* 2006) and response to therapy studies (Krak *et al* 2005) are rapidly gaining ground. Whereas accurate activity concentration recovery is crucial for correct diagnosis and monitoring response to therapy, applications such as the use of PET in intensity-modulated radiation therapy (IMRT) treatment planning render equally vital the accurate shape and volume determination of lesions. Different volume-of-interest (VOI) determination methodologies have been proposed that can be classified as manual or automatic. On the one hand, manual segmentation methods to delineate boundaries are laborious and highly subjective (Krak *et al* 2005). On the other hand, automatic segmentation of objects of interest in PET (Reutter *et al* 1997, Zhu and Jiang 2003, Kim *et al* 2002, Riddell *et al* 1999) is not a trivial task because of low spatial resolution and resulting partial volume effects (PVE), low contrast ratios, as well as noise resulting from the statistical nature of radioactive decay or the choice of the reconstruction process.

The most widely used method to semi-automatically determine VOIs in PET is thresholding, either adaptive, using *a priori* computed tomography (CT) knowledge (Erdi *et al* 1997), or fixed threshold (Krak *et al* 2005) using values derived from phantom studies (from 30 to 75% of maximum local activity concentration value) (Jarritt *et al* 2006, Krak *et al* 2005, Erdi *et al* 1997). Such thresholding techniques, however, lead to variable VOI determination as shown in recent clinical studies (Nestle *et al* 2005). On the other hand, numerous works have addressed automatic lesion detection from PET datasets, including different methodologies such as edge detection (Reutter *et al* 1997), fuzzy C-means (Zhu and Jiang 2003), clustering (Kim *et al* 2002) or watersheds (Riddell *et al* 1999). The performance of these algorithms is sensitive to variations of noise intensity and/or lesion contrast. In addition, past work has in its majority considered the ability of such automatic methodologies for the detection of lesions but not the accuracy with which the methods are capable for VOI and/or activity concentration determination. Furthermore, all of the afore-mentioned algorithms often involve user-dependent initializations, pre- and post-processing, or additional information like CT or expert knowledge; rendering their employment more complicated and the outcome dependent on choices made by the user in relation to the pre- and/or post-processing steps necessary. For example in the case of the watershed algorithm a filtering pass as a pre-processing step to smooth the image, and a post-processing step to fuse the different regions resulting from the algorithm are necessary.

Hidden Markov models are automatic segmentation algorithms allowing noise modelling and have proven to be less sensitive to variation of the values in the regions of the images than other segmentation approaches thanks to their statistical modelling. They have only been previously used in PET in the form of hidden Markov fields (HMF) (Chen *et al* 2001). Hidden Markov chains (HMC) (Benmiloud and Pieczynski 1995) is a faster model and can offer competitive results (Salzenstein and Pieczynski 1998). Furthermore, HMC leads to shorter computational times, as quantities of interest can be computed directly on the chain, whereas

the HMF algorithm needs iterative Monte Carlo like estimation procedures (Salzenstein and Pieczynski 1998) that are time consuming. These algorithms offer an unsupervised estimation of the parameters needed for the image segmentation and limit the user's input to the number of classes to be searched for in the image. Reconstructed images require no further pre- or post-processing treatment (such as for example filtering) prior to the segmentation process. Instead, image noise is considered as additional information (a parameter in the classification decision process) to be taken into account, not to be suppressed or avoided.

The objectives of our study were to (a) develop a new fuzzy HMC (FHMC) model in an attempt to account for the limited spatial resolution in PET and (b) compare the performance of FHMC with those of the thresholding methodologies currently used in clinical practice. Different imaging conditions in terms of statistical quality, as well as lesion size and source-to-background (S/B) ratio were considered in this study. The analysis was carried out on both simulated and acquired images reconstructed using iterative algorithms which form today's state of the art in whole body PET imaging in routine clinical oncology practice (Visvikis *et al* 2001, 2004).

## 2. Materials and methods

### *2.1. Hard and fuzzy hidden Markov chain models*

The HMC model is an unsupervised methodology that takes place in the Bayesian framework. Although we place ourselves in the application of image segmentation this methodology can be used in other applications such as, for example, speech recognition (Dai 1994). Let $T$ be a finite set corresponding to the voxels of an image. We consider two random processes $\mathrm{Y} = (y_t)_{t\in T}$ and $\mathrm{X} = (x_t)_{t\in T}$. Y represents the observed image, and X represents the 'hidden' segmentation map. X takes its values in $\Omega = \{1, \ldots, K\}$ with $K$ being the number of classes $c$, and Y takes its values in $\mathbb{R}$. We assume that a Markov process can model the prior distribution of X. The segmentation problem consists in estimating the hidden X from the available noisy observation Y. The relationship between X and Ycan be modelled by the joint distribution $P(\mathrm{X}, \mathrm{Y})$. This distribution can be obtained thanks to the Bayes formula:

$$P(\mathrm{X}|\mathrm{Y}) = \frac{P(\mathrm{X}, \mathrm{Y})}{P(\mathrm{Y})} = \frac{P(\mathrm{Y}|\mathrm{X})P(\mathrm{X})}{P(\mathrm{Y})}. \tag{1}$$

$P(\mathrm{Y}|\mathrm{X})$ is the likelihood of the observation Y conditionally with respect to the hidden ground-truth X, and $P(\mathrm{X})$ is the prior knowledge concerning X. The Bayes rule allows us to know the posterior distribution of X with respect to the observation Y. In the Markov chain framework we have to assume the random variables $\mathrm{Y} = (y_t)_{t\in T}$ are conditionally independent with respect to X and that the distribution of each $y_t$ conditional on X is equal to its distribution conditional on $x_t$. Many applications of hidden Markov models with unsupervised estimation have been successful considering different types of images (radar, sonar, magnetic resonance images (MRI), CT, satellite or astronomical) (Pieczynski 2003, Salzenstein *et al* 2004, Delignon *et al* 1997), but this kind of approach was almost never applied to PET data.

*2.1.1. Markov chain definition.* X is a Markov chain if:

$$P(x_t|x_1, \ldots, x_{t-1}) = P(x_t|x_{t-1}) \qquad \text{for} \quad 1 < t \leqslant T. \tag{2}$$

The distribution of X is then defined by the distribution of $x_1$, called initial probabilities init($c$) for each class $c$ ($P(x_1 = c)$) and the transition matrix trans($c, d$) (of dimension $K \times K$) containing the probabilities of transitions from the class $c$ to the class $d$; $P(x_{t+1} = d|x_t = c)$. As X and Y are one-dimensional elements in the HMC context, a spatial transformation

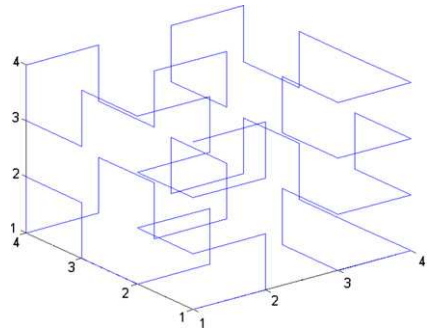



**Figure 1.** The 3D Hilbert–Peano space filling curve for a 4 × 4 × 4 voxels VOI.

is necessary to process three-dimensional VOIs. For the best preservation of the spatial correlation between voxels we use the Hilbert–Peano space-filling curve. This fractal path can be extended to explore 3D VOIs (Kamata *et al* 1999). A visual illustration of the Hilbert–Peano path for a 4 × 4 × 4 voxels 3D VOI is given in figure 1. Once the chain has been segmented, the inverse path is used to reconstruct the 3D segmentation map.

*2.1.2. Adding a fuzzy measure to the model.* The general idea behind the implementation of a fuzzy model within the Bayesian framework was previously introduced by Salzenstein (Salzenstein and Pieczynski 1997). Its implementation in association with HMC developed as a part of this work is based on the incorporation of a finite number of fuzzy levels $F_i$ in combination with two homogeneous (or 'hard') classes, in comparison to HMC where only a finite number of hard classes are considered. This model allows the coexistence of voxels belonging to one of two hard classes and voxels belonging to a 'fuzzy level' depending on its membership to the two hard classes. Therefore, FHMC adds an estimation of imprecision of the hidden data (X, see section 2.1) in contrast to HMC which only models uncertainty of the observed data (Y, see section 2.1). The statistical part of the algorithm models the uncertainty of the classification, with the assumption being that the voxel is clearly identified but the observed data is noisy. On the other hand, the fuzzy part models the imprecision of the voxel's membership, with the assumption being that the voxel may contain both classes. One way to achieve this extension is to simultaneously use Dirac and Lesbegue measures at the class level. Hence we consider that X in the fuzzy model takes its values in [0, 1] instead of $\Omega = \{1, \ldots, K\}$. Let $\delta_0$ and $\delta_1$ be the Dirac measures at 0 and 1, and $\zeta$ the Lesbegue measure on ]0, 1[. We define the new measure $\nu = \delta_0 + \delta_1 + \zeta$ on [0, 1]. Note that, for example, using two hard classes and two fuzzy levels in the FHMC model is not equivalent to using four hard classes in the HMC model where $\nu = \delta_1 + \delta_2 + \cdots + \delta_K$. This has been previously stated using Markov fields based segmentation (Salzenstein and Pieczynski 1997).

The distribution of X can then be defined using a conjoint density $g$ for $(x_t, x_{t+1})$ on [0, 1] × [0, 1]:

$$\begin{aligned} &\text{let } (a, b) \in [0, 1] \times [0, 1] \\ &g(a = 0, b = 0) = \alpha_1 \qquad \text{and} \qquad g(a = 1, b = 1) = \alpha_2 \\ &g(a = 0, b = 1) = \gamma_1 \qquad \text{and} \qquad g(a = 1, b = 0) = \gamma_2 \\ &g(a, b) = \beta f_g(a, b) \qquad \text{if} \quad (a, b) \neq \{(0, 0), (0, 1), (1, 0), (1, 1)\} \end{aligned} \tag{3}$$

with

$$\int_{[0,1]} \int_{[0,1]} g(a, b)\, \mathrm{d}(\nu \otimes \nu)(a, b) = 1 \qquad \text{and} \qquad \alpha_1 + \alpha_2 + \gamma_1 + \gamma_2 + \beta\lambda = 1 \tag{4}$$

where $\mathrm{d}(\nu \otimes \nu)(a, b)$ is the notation for integration with respect to the $(a, b)$ variables, each one being with respect to the measure $\nu$ on the interval $[0, 1]$. $\lambda$ is a constant depending on the form of the parameterized function $f_g$:

$$f_g(a, b) = 1 - |a - b|. \tag{5}$$

We now define the initial and transition probabilities (init$(c)$ and trans$(c, d)$) using the conjoint density $g$ and an utility density $h$ on $[0, 1]$ defined by: $h(a) = \int_0^1 g(a, b)\,\mathrm{d}\nu(b)$:

init$(c)$ using densities $g$ and $h$:

$$\begin{aligned} P\,(x_1 \in \{0, 1\}) &= \int_0^1 g(x_1, b)\,\mathrm{d}\nu(b) = h(x_1) \\ P\,(x_1 \in F_i) &= \int_{\frac{i-1}{N}}^{\frac{i}{N}} \int_0^1 g(a, b)\,\mathrm{d}(\nu \otimes \nu)(a, b) \simeq \frac{1}{N} \int_0^1 g(\varepsilon_i, b)\,\mathrm{d}(\nu)(b) = \frac{1}{N} h(\varepsilon_i) \end{aligned} \tag{6}$$

trans$(c, d)$ using the conditional density $f$ deduced from (1) : $f(x_{t+1}|x_t) = \frac{g(x_{t+1}, x_t)}{h(x_t)}$

$$\begin{aligned} P(x_{t+1} \in F_j | x_t \in F_i) &= \frac{\int_{F_j} \int_{F_i} g(\varepsilon_j, \varepsilon_i)\,\mathrm{d}(\nu \otimes \nu)(\varepsilon_j, \varepsilon_i)}{\int_{F_i} h(\varepsilon_i)\,\mathrm{d}\nu(\varepsilon_i)} \\ P(x_{t+1} \in F_j | x_t \in \{0, 1\}) &= \frac{\int_{F_j} g(\varepsilon_j, x_t)\,\mathrm{d}\nu(\varepsilon_j)}{h(x_t)} \\ P(x_{t+1} \in \{0, 1\} | x_t \in F_i) &= \frac{\int_{F_i} g(x_{t+1}, \varepsilon_i)\,\mathrm{d}\nu(\varepsilon_i)}{\int_{F_i} h(\varepsilon_i)\,\mathrm{d}\nu(\varepsilon_i)} \\ P(x_{t+1} \in \{0, 1\} | x_t \in \{0, 1\}) &= \frac{g(x_{t+1}, x_t)}{h(x_t)} \end{aligned} \tag{7}$$

where $N - 1$ is the number of fuzzy levels and $\varepsilon_i = \frac{i}{N}$ is the value associated with a fuzzy level $F_i$.

The fuzzy model is a generalization of the hard model. The use of the Dirac measures allows one to retrieve the standard two-class hard model when the fuzzy component is null. As the theoretical framework described above has not been developed for a specific kind of image, but as a general segmentation algorithm, the *a priori* and the noise (also called observation) models are not directly derived from PET image characteristics. However this segmentation approach may be appropriate in segmenting PET images since they are both noisy and of low resolution. The 'noise' aspect when considering hidden Markov models in general is the way the values of each class to be found in the image are distributed around a mean value. The noise model used, whose respective mean and variance are to be determined by the estimation steps, can therefore be adapted to image specific characteristics. On the other hand, the fuzzy measure allows a more realistic modelling of the objects' border transitions between foreground and background, allowing in such a way to indirectly account for the effects of blurring (partial volume effects) associated with low resolution images, such as those in PET.

*2.1.3. Segmentation and parameters estimation.* In order to perform segmentation on the chain level, we need to use a criterion to classify each element as background or functional VOI. For this purpose we use the marginal posterior mode (MPM) (Marroquin *et al* 1987). This approach aims to minimize the expectation $E\{L(x_t, \widehat{x}_t)|Y\}$ where $L$ is a loss (or cost) function:

$$L(x_t, \hat{x}_t) = |x_t - \hat{x}_t| \tag{8}$$

with $x_t$ the real class and $\hat{x}_t = \hat{s}(y_t)$ the one affected by the segmentation process $\hat{s}$. This criterion is adequate for the segmentation problem as it penalizes a configuration with respect to the number of misclassified elements. In order to compute a solution, the MPM segmentation needs the parameters defining the *a priori* model (initial and transition probabilities of the chain) as well as the noisy observation data model (mean and variance of each class). The assumption that the noise for each class of the observed data can fit a Gaussian distribution was made as a first step. The mean and variance of each fuzzy level $F_i$ is derived from the ones estimated in the two hard classes as follows:

$$\mu_{F_i} = \mu_0(1-\varepsilon_i) + \varepsilon_i\mu_1 \qquad \sigma^2_{F_i} = \sigma^2_0(1-\varepsilon_i)^2 + \varepsilon_i^2\sigma^2_1. \tag{9}$$

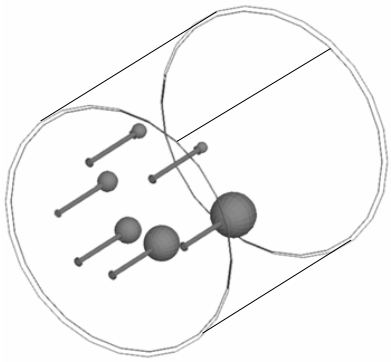
Both *a priori* and noise models parameters are unknown in the real case and therefore they must be estimated. In order to achieve such estimation, we use the stochastic iterative procedure called stochastic expectation maximization (SEM) (Celeux and Diebolt 1986), a stochastic version of the EM algorithm (Dempster *et al* 1977). This is achieved in a similar fashion to that used in the classical HMC context by simulating posterior realizations of X (see the appendix for detailed posterior realization of X and the SEM procedure) and computing empirical values of the parameters of interest using the simulated chain. The stochastic nature of this procedure makes it less sensitive to the initial guess of the parameters using the K-means (McQueen 1967) than deterministic procedures such as the EM algorithm. Both the MPM segmentation and SEM parameters estimation use a practical recursive computation of the values of interest called forward–backward procedure that is performed directly on the chain (Benmiloud and Pieczynski 1995). The implementation of the FHMC segmentation algorithm in a step-by-step fashion can be found in the appendix. Note that the overall algorithm is entirely unsupervised (except for the number of classes and fuzzy levels to use) and it is able to adjust to a large spectrum of image structures, noise or contrast. For example, no *a priori* is made on the shape of the objects to extract or the source-to-background ratio in the image.

### *2.2. Thresholding*

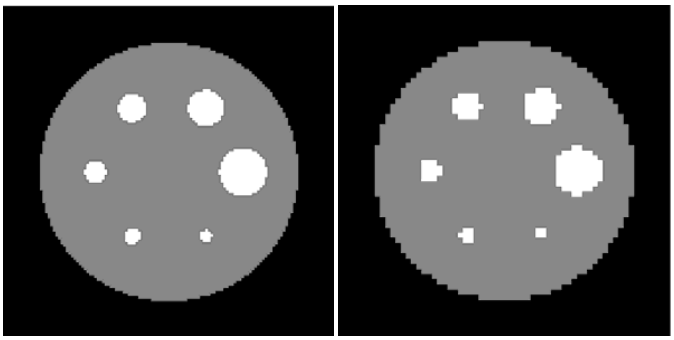
Various thresholding methodologies have been proposed in the past for both functional volume segmentation and/or activity concentration recovery (Krak *et al* 2005, Erdi *et al* 1997, Nestle *et al* 2005). Thresholding using 42% and 50% of the maximum value in the lesion was chosen for VOI determination and quantitation purposes respectively, based on previous publications (Krak *et al* 2005, Erdi *et al* 1997). The methodology was implemented through region growing using the voxel of maximum intensity in the object of interest as a seed. Using a 3D neighbourhood (26 neighbours) the region is iteratively increased by adding neighbouring voxels if their intensity is superior or equal to the selected threshold value. The results derived using these methods will be denoted from here onwards as T42 and T50 for the thresholds of 42% and 50% respectively.

### *2.3. Validation studies*

*2.3.1. Simulated and acquired datasets.* Simulated datasets using the IEC image quality phantom (IEC 1998), containing six different spherical lesions of 10, 13, 17, 22, 28 and 37 mm in diameter (figure 2) were generated using Geant4 Application for Tomographic Emission (GATE) and a validated model of the Philips Allegro PET scanner (Lamare *et al* 2006). Images, considering only the detected true coincidences, were subsequently reconstructed using the OPL-EM iterative algorithm (Reader *et al* 2002) with seven iterations (Lamare *et al* 2006). Two different voxel sizes were considered in the reconstructed images;



(a)



(b) (c)

**Figure 2.** (a) A graphical representation of the IEC phantom, and the central slice of the digital IEC phantom used in the generation of the simulated datasets (b) with $2 \times 2 \times 2$ mm$^3$ and (c) $4 \times 4 \times 4$ mm$^3$.

namely $2 \times 2 \times 2$ mm$^3$ and $4 \times 4 \times 4$ mm$^3$. The 8 mm$^3$ voxel size configuration leads to better sampled objects of interest but with higher noise due to the number of counts being divided by eight in each voxel in comparison to the 64 mm$^3$ voxel sizes. A uniform activity was simulated throughout the phantom cylinder and the lesions. Different parameters were however considered to cover a large spectrum of configurations allowing assessment of the influence of different parameters susceptible to affect the functional VOI determination or quantitation accuracy. The statistical quality of the images was varied by considering 20, 40 and 60 million true coincidences. Two different signal-to-background (S/B) ratios were also considered, 4:1 and 8:1 (with around 6 kBq cm$^{-3}$ in the background, and 24 or 48 kBq cm$^{-3}$ in the spheres respectively). Visual illustration of the reconstructed images corresponding to different simulated configurations is given in figures 3(a)–(d).

In addition to the simulated datasets, acquisitions of the IEC phantom were carried out in the list-mode format using a Philips GEMINI PET/CT scanner. The only difference with the simulated datasets was the exclusion of the 28 mm diameter sphere in the study because in the phantom used it was replaced by a plastic sphere of unknown diameter. The same S/B ratios of 4:1 and 8:1 used in the simulations were also employed in this part of the study, by introducing 7.4 kBq cm$^{-3}$ in the background and 29.6 or 59.2 kBq cm$^{-3}$ respectively in the spheres. Different count statistical qualities were obtained by reconstructing 1 min, 2 min or 5 min list-mode time frames using the 3D RAMLA algorithm, with specific parameters

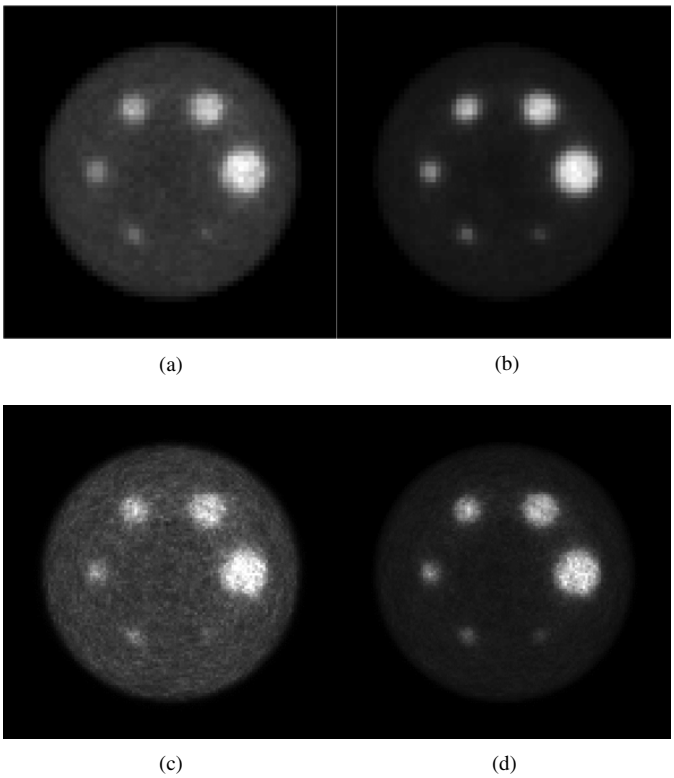
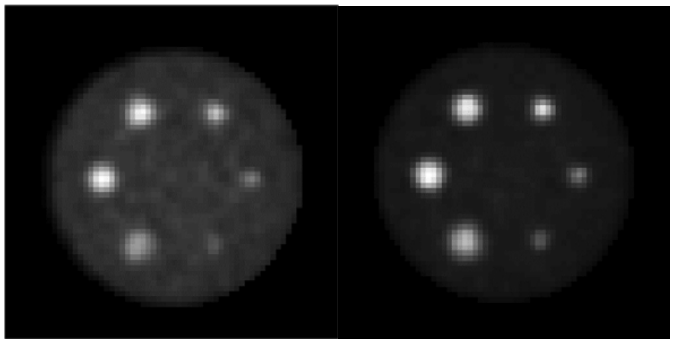


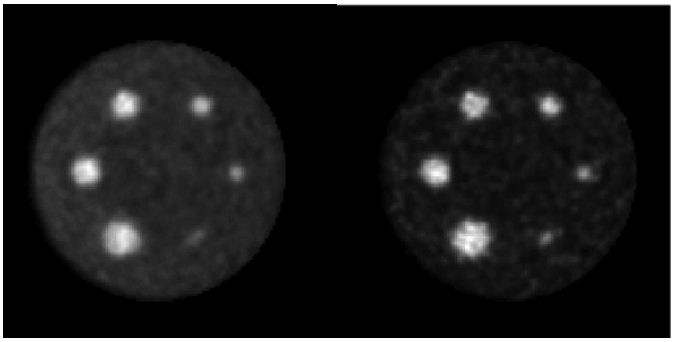
**Figure 3.** Different images used in the segmentation study; (a)–(d) simulated: (a) ratio 4:1, 20 million coincidences, 64 $mm^3$, (b) ratio 8:1, 40 million, 64 $mm^3$, (c) ratio 4:1, 20 million coincidences, 8 $mm^3$, (d) ratio 8:1, 40 million, 8 $mm^3$; (e)–(h) acquired: (e) ratio 4:1, 2 min acquisition time, 64 $mm^3$, (f) ratio 8:1, 5 min, 64 $mm^3$, (g) ratio 4:1, 5 min, 8 $mm^3$, (h) ratio 8:1, 5 min, 8 $mm^3$.

previously optimized (Visvikis *et al* 2004). The same voxel sizes as for the simulated datasets (8 $mm^3$ and 64 $mm^3$) were used in the reconstruction of each of the different statistical quality datasets considered. Visual illustration of the acquired images is given in figures 3(e)–(h). Each sphere in both simulated and acquired images was isolated in a box of the same size ($16 \times 16 \times 10$ for the 4 mm case, and $32 \times 32 \times 20$ for the 2 mm case) prior to the segmentation process.

*2.3.2. Computed volume versus classification error measurement.* The majority of previous works dealing with VOI determination in PET measure the performance of a given methodology by computing the VOI obtained on the segmentation map and comparing it with the true known volume of the object of interest. This type of approach has the potential to lead to biased performance measurements since a segmentation result may contain two different types of errors. On the one hand, one may have voxels of the background that are classified as belonging to the object of interest, denoted from here on as positive classification errors (PCE), while on the other hand, one may end up with voxels of the object that are classified as belonging to the background, denoted from here on as negative classification errors (NCE). These classification errors essentially occur on the boundaries of the objects of interest because of 'spill in' (increasing probabilities of a NCE) and 'spill out' (increasing probabilities of a



(e) (f)



(g) (h)

**Figure 3.** (Continued.)

PCE). If the segmentation results in PCEs and NCEs of equal amounts, the computed VOI would be very close to the true known volume whereas the shape and position of the object would be incorrect. The shape and position information is as important as the total volume of the object in order to accurately derive a radiotherapy treatment planning or the activity concentration of interest in a response to therapy study based on the derived functional volume. For example, let us assume that the segmentation process results in 20% NCEs and 15% PCEs. This leads to a classification error of 35% whereas the error in the overall computed volume is only $-5\%$. Hence, the use of classification error is a more pertinent measurement of the accuracy with which a given algorithm performs the task of functional volume delineation since it takes into account not only the segmented volume in comparison to the actual volume of interest but also its position and shape.

In the simulation study the total number of PCEs and NCEs is considered with respect to the number of voxels defining the sphere (VoS) in the digital phantom (the ground truth) in order to obtain a percentage classification error (CE):

$$\mathrm{CE} = \frac{(\mathrm{PCE} + \mathrm{NCE})}{\mathrm{VoS}} \times 100. \tag{10}$$

The size of classification errors can be bigger than 100% in the case where a large number of background voxels in the selected area of interest are misclassified as belonging to the sphere. In practical terms, maximum classification errors calculated during this work where limited to 200%, since any such values represent complete failure of the segmentation process. In addition, the interest of classification errors is when they occur at the borders of the objects

and not in other regions of the background. One should also keep in mind that a combined representation of PCE and NCE into CE leads to a loss of information as far as the direction of the bias is concerned. It does, however, still represent more pertinent information than overall volume estimation errors, which reflect neither accurate magnitude nor direction of the bias for a segmented volume.

On the other hand in the case of the images reconstructed from the acquired datasets only overall computed volumes were considered in order to avoid any biases as a result of misalignment and rescaling inaccuracies, as well as reconstruction artefacts in the higher and lower slices of the associated CT datasets. As the goal is not to detect the lesion in the whole image but to estimate its volume, shape and position with the best accuracy possible, we assume that the lesion has been previously identified by the clinician and automatically or manually placed in a 3D 'box' well encompassing the object. Subsequently, the images of the selected area were segmented in two classes (*functional VOI* and *background*) using each of the three methods under evaluation (thresholding, FHMC and HMC). In the FHMC case, different numbers of fuzzy levels were considered in the segmentation process (namely two and three). Following the segmentation by FHMC, volumes of interest can be defined using the hard classes and any number of the fuzzy levels considered.

*2.3.3. Quantitation accuracy.* In terms of quantitation the objective of our study was to determine the accuracy of the average activity concentration recovered from a volume derived using a given segmentation algorithm. The 'ground truth' for comparison purposes was established using the exact size, shape and location of each lesion (using the known digital phantom employed in the generation of the simulated datasets).

As a result, these recovered activity concentration values represented an under-estimation of the true activity due to PVE. A comparison on a lesion-by-lesion basis was subsequently carried out with the measured activity concentration from the segmented volumes obtained by the three algorithms considered. T50 should lead to some improvements in the lesion activity recovery with respect to T42 as a result of including less voxels in the volume used to compute the activity and therefore less voxels associated with PVE. Similarly FHMC $0/2$ (see section 3, results, for the definition of FHMC $x/y$) should lead to concentration recovery improvements with respect to FHMC $1/2$, since voxels belonging to the fuzzy levels are found at the edges of the lesions and their intensity is most significantly reduced by PVE. Therefore the inclusion of these voxels should only result in even stronger under-evaluation of the true lesion activity concentrations.

## 3. Results

Different segmentation maps obtained using each of the methods under evaluation are presented in figure 4 for a slice centred on the 28 mm sphere of the simulated images to visually illustrate the variations of the segmentation maps obtained. Figure 5(a) shows the impact of the number of fuzzy levels included in the FHMC segmentation. The various FHMC maps are denoted as FHMC $x/y$ with $x$ being the number of fuzzy levels included in the segmentation map, and $y$ being the total number of fuzzy levels used in the segmentation process. The error bars in these figures represent different results obtained for each of the three different levels of statistical quality considered (the top of the error bar is the result concerning the worst statistical quality, the medium one concerns the medium quality and the lowest one corresponds to the best quality considered). As figure 5(a) shows, for the range of simulated spheres considered, no improvement was obtained in the % classification errors by having more

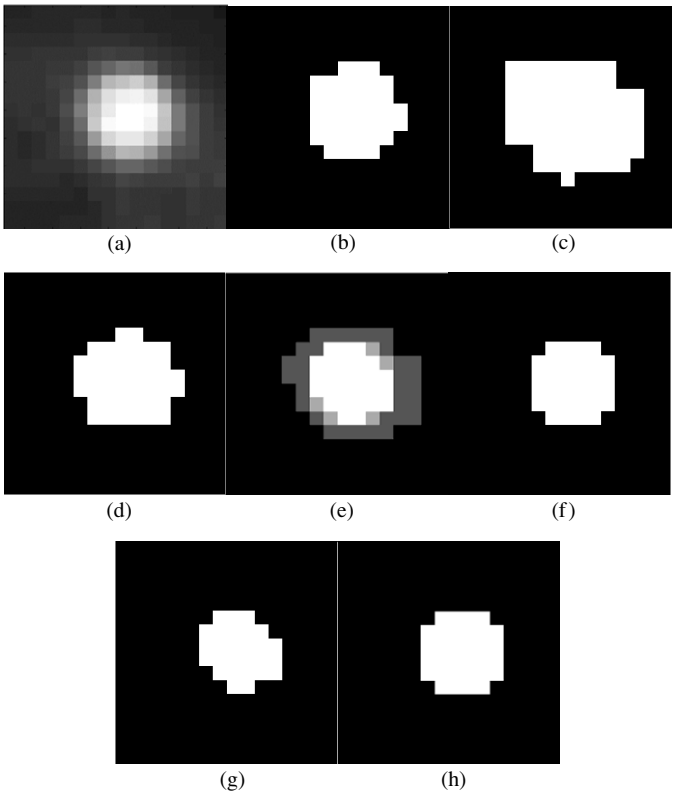
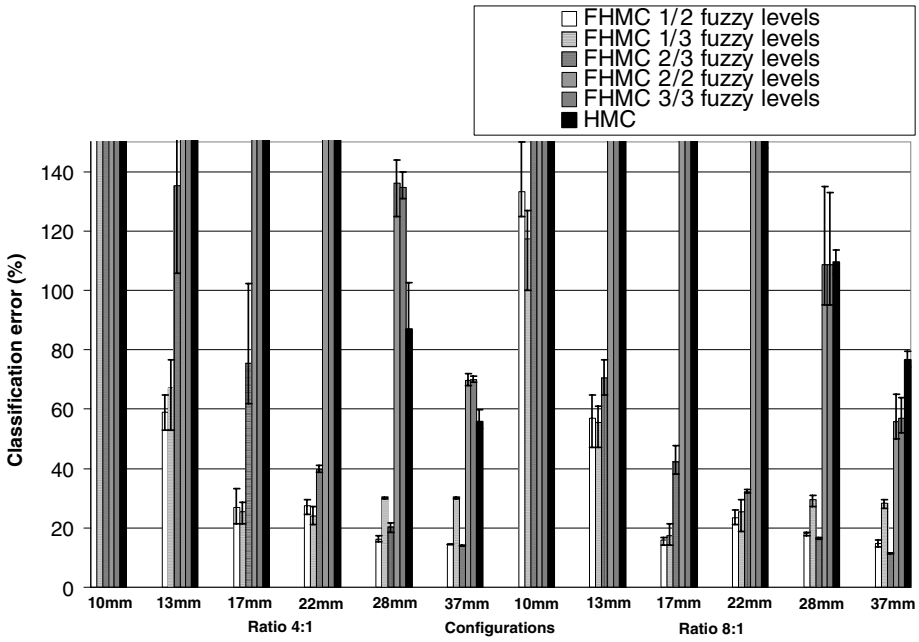


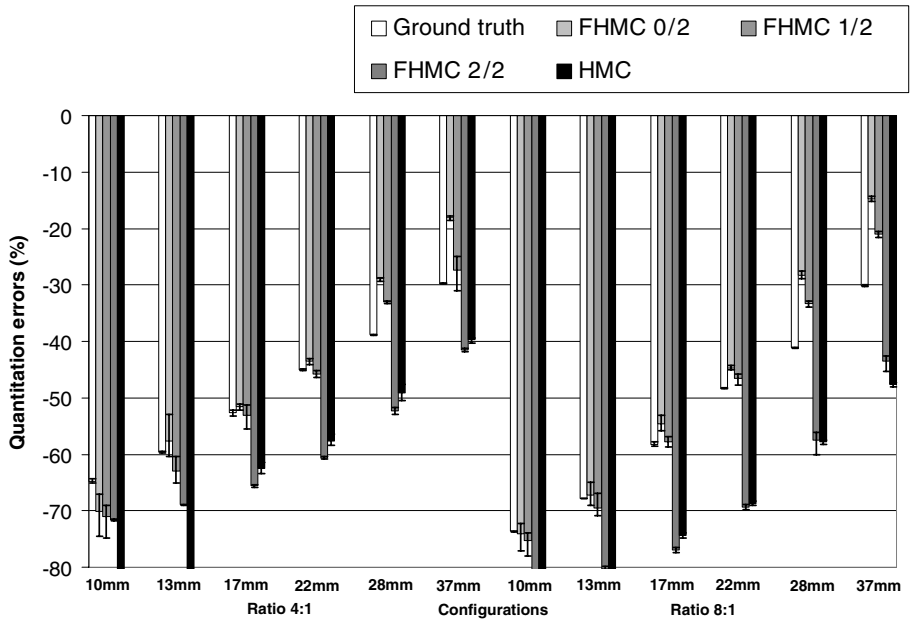

**Figure 4.** Examples of segmentation maps for the 28 mm sphere (one slice): (a) PET ROI, (b) digital 'ground truth', (c) HMC map, (d) T42 map, (e) FHMC with two fuzzy levels (light and dark grey voxels) segmentation map, (f) map used for VOI (hard class + first fuzzy level, FHMC 1/2), (g) map for quantitation (only hard class voxels, FHMC 0/2), (h) T50 map. Note that in this particular case, FHMC 1/2 for VOI and T50 result in the same map but this is of course not always the case (especially having considered the complete 3D volume).

than two fuzzy levels in the FHMC segmentation process and keeping in the overall segmented volume more than the voxels identified in the first fuzzy level. It should be emphasized at this point that this conclusion was reached considering the results of the whole of the range of simulated sphere diameters and keeping in mind that our objective is determining a single best configuration of the algorithm parameters across a wide range of imaging conditions and not different parameters for individual lesion sizes, image statistics or contrast ratios. In addition, it is clearly shown in figure 5 that HMC leads to worse segmentation results in comparison to FHMC for all different configurations considered. Therefore for all subsequent volume determination analyses, the results associated with the FHMC 1/2 versus T42 are presented. As shown in figure 5(b), no benefits are observed through the inclusion in the segmentation map of any voxels belonging to the fuzzy domain. This confirms what was anticipated in section 2.3.3. Therefore from here onwards all the quantitation results presented for FHMC have been calculated using only the hard class voxels resulting from the segmentation process (FHMC 0/2).

The % classification errors for reconstructed images of the simulated datasets as a function of lesion size and contrast are presented in figure 6(a) for 64 $mm^3$ and (b) for 8 $mm^3$, for the FHMC and the threshold based method (T42). A breakdown, in terms of PCEs




(a)




(b)

**Figure 5.** (a) Classification errors for the lesion VOI determination and (b) lesion activity recovery errors in the simulated images for the FHMC versus HMC segmentation. Different numbers of fuzzy levels (two or three) were used in the segmentation process and different numbers of these were subsequently selected to (a) form the segmented volumes or (b) determine lesion average activity concentrations for the different imaging conditions considered.

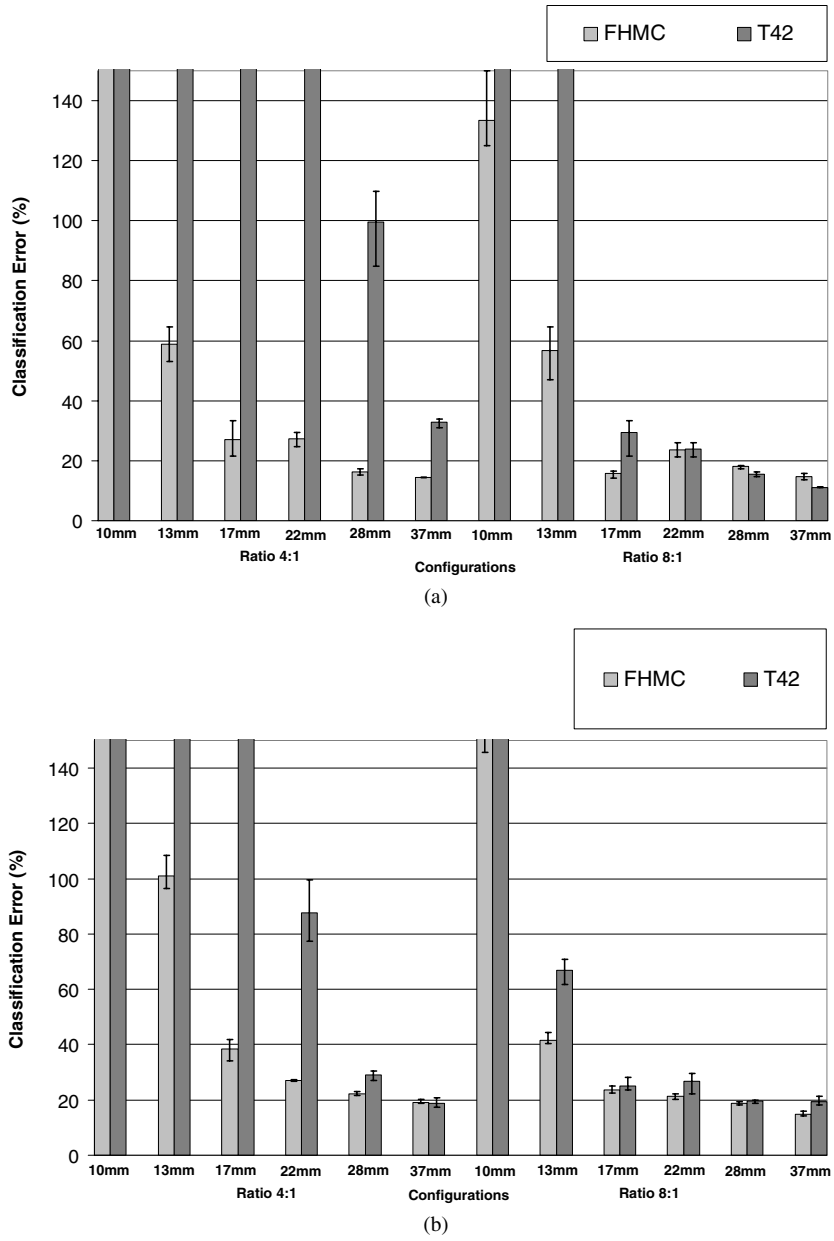



**Figure 6.** Classification errors in lesion VOI determination from the simulated images: (a) 64 $mm^3$ voxels and (b) 8 $mm^3$ voxels for the FHMC 1/2 versus T42 segmentation under variable imaging conditions.

and NCEs, of the % classification errors in figure 6(a) is given in figures 7(a)–(c) for the FHMC, HMC and T42 segmentation methods respectively. Finally, in order to facilitate a comparison of the segmentation results between the simulated and the acquired datasets, the % computed volume error is given in figures 8(a)–(b) for the same configurations as in figures 6(a)–(b).

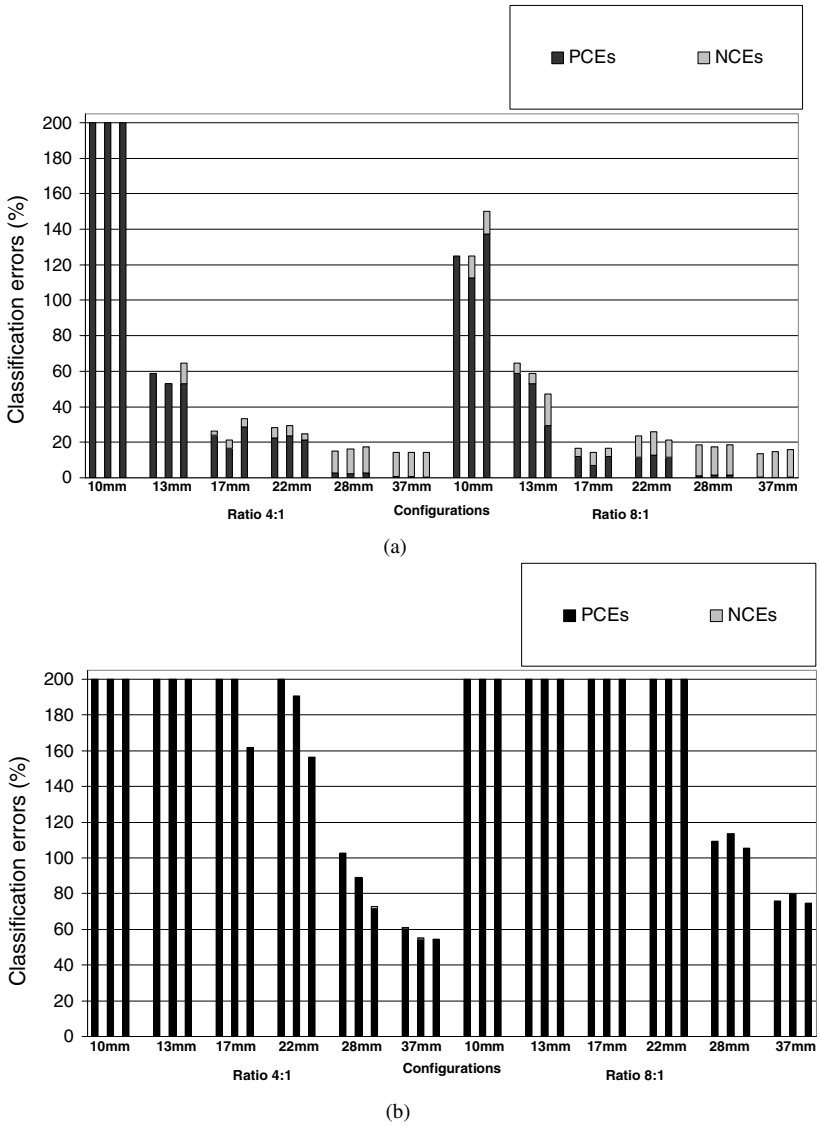
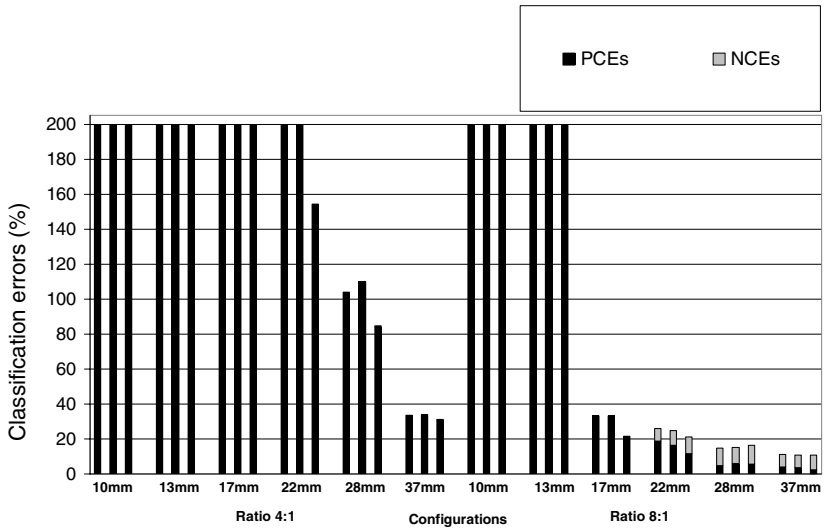



**Figure 7.** Repartition of PCEs and NCEs from the (a) FHMC 1/2, (b) HMC and (c) T42 segmentation results for the different simulated imaging configurations considered.

Considering the simulated datasets, the introduction of FHMC led to superior results in comparison to the current 'gold standard' in functional volume delineation of T42. FHMC segmentations led to $<25\%$ classification errors in computed volumes for lesion sizes $>13$ mm irrespective of contrast ratio, level of noise or lesion size. Errors of more than 200% for FHMC were only observed for the 10 mm sphere. Results for the T42 were more dependent on the lesion size, relative to FHMC results, varying from 10% to more than 200% (even for spheres up to 22 mm in diameter for a contrast of 4:1 and 64 $mm^3$ voxel size). However, the use of T42 was found to work well for lesion sizes of $>17$ mm and a lesion-to-background ratio of 8:1 with % classification errors of 20–30%. On the other hand, for a lesion-to-background ratio of 4:1, the T42 threshold led to over 100% overestimation in the functional volume

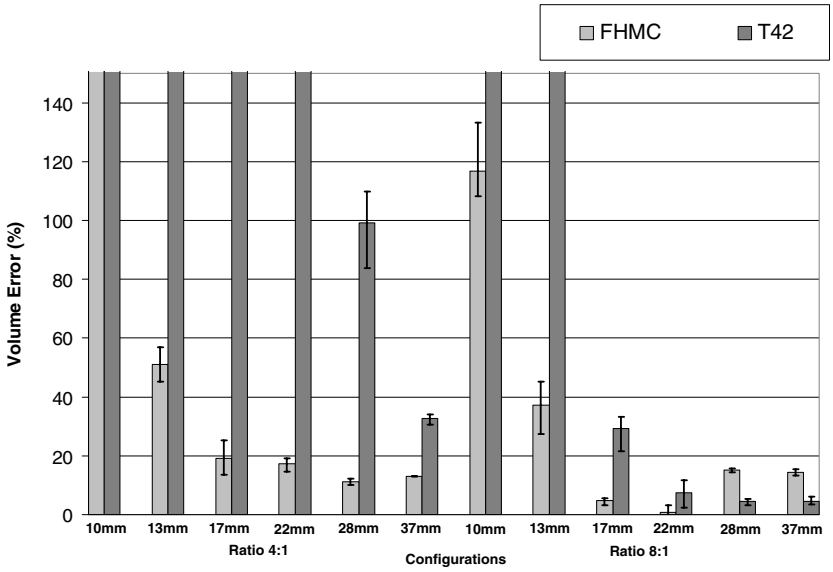



(c)

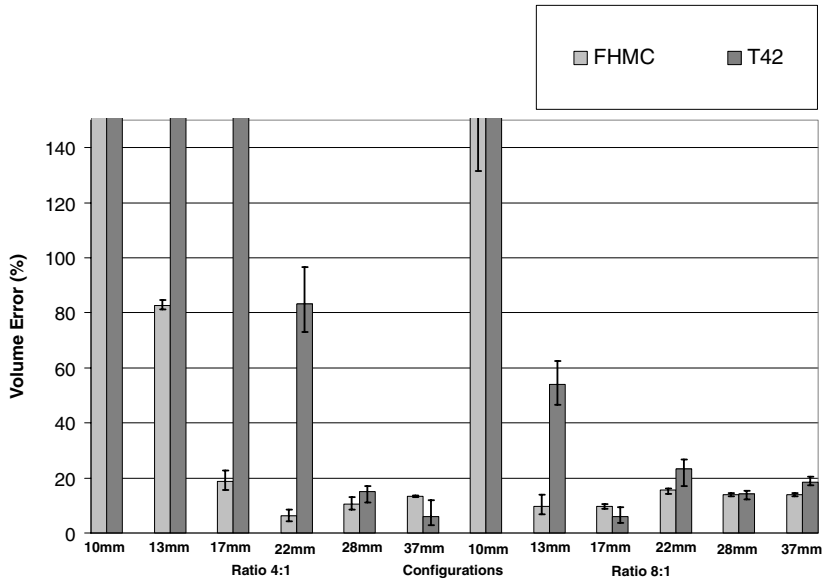
**Figure 7.** (Continued.)

for lesions <28 mm in diameter. As the errors bars in the different figures reveal, there was a larger dependence on the statistical quality of the reconstructed images observed with T42 in comparison to FHMC for the majority of the lesion sizes and contrast configurations considered. In particular this was true for all of the lesions for a contrast ratio of 4:1 and for lesions <22 mm for a contrast ratio of 8:1. For example, for the 17 mm sphere and a contrast ratio of 8:1, T42 resulted in classification errors of 20 to 35% whereas FHMC classification errors from 15 to 17% were observed (figure 6). On the other hand in the case of the 28 mm sphere and a contrast ratio of 4:1, T42 errors were ranging from 85 to 110% whereas FHMC resulted in errors of 17–18%. The reduction in the reconstruction voxel size (from 64 $mm^3$ to 8 $mm^3$) led to small differences in the functional volumes determined using the FHMC segmentation algorithm, and although it led to improvements in the T42 based segmented volumes, the % classification errors remained at 80–200%. The trend observed with the standard voxel sizes on the variation of the segmentation results as a function of statistical quality was similar for the reduced voxel size images. For example in the case of the 22 m sphere and a contrast ratio of 4:1 errors of 77–100% and 26–27% were observed for T42 and FHMC respectively. In general, the largest errors were observed for the smaller lesions of 10 and 13 mm, where none of the segmentation algorithms considered performed well under any of the configurations tested, with errors largely >200%. As shown in figure 7(a) FHMC classification errors are essentially NCEs for the two biggest spheres and PCEs for the small ones. In contrast, as shown in figures 7(b)–(c), T42 and HMC methods result essentially in PCEs, apart from T42 in association with lesions >28 mm in diameter and a lesion-to-background ratio of 8:1.

In terms of overall volume estimation errors on simulated datasets (see figures 8(a), (b)) FHMC results in errors of up to 10% and between 10% and 20% for a contrast ratio of 8:1 and 4:1 respectively, for lesions >13 mm. T42 led to volume determination errors of <10% for lesions >17 mm in diameter and a lesion-to-background ratio of 8:1, while errors of over 100% were observed for lesions <28 mm with a lesion-to-background ratio of 4:1. However, while the lowest overall volume error of T42 was around 10%, the corresponding classification error was >20%. In the case of an 8 $mm^3$ reconstructed voxel size (figure 8(b)) small improvements




(a)




(b)

**Figure 8.** Lesion VOI estimation errors from the simulated images: (a) 64 $mm^3$ voxels and (b) 8 $mm^3$ voxels for the FHMC 1/2 versus T42 segmentation under variable imaging conditions.

were seen using the T42 for lesions $\geqslant$13 mm and $>$22 mm for a lesion-to-background ratio of 8:1 and 4:1 respectively. Finally, no noticeable differences were seen in the FHMC based segmentation results, apart from an improvement to $<$15% in the volume estimation error for the 13 mm lesion with a contrast size of 8:1.

Figures 9(a), (b) show the results in terms of % error in the recovered activity as a function of lesion size and contrast ratio considering the segmented volumes using 64 $mm^3$ and 8 $mm^3$

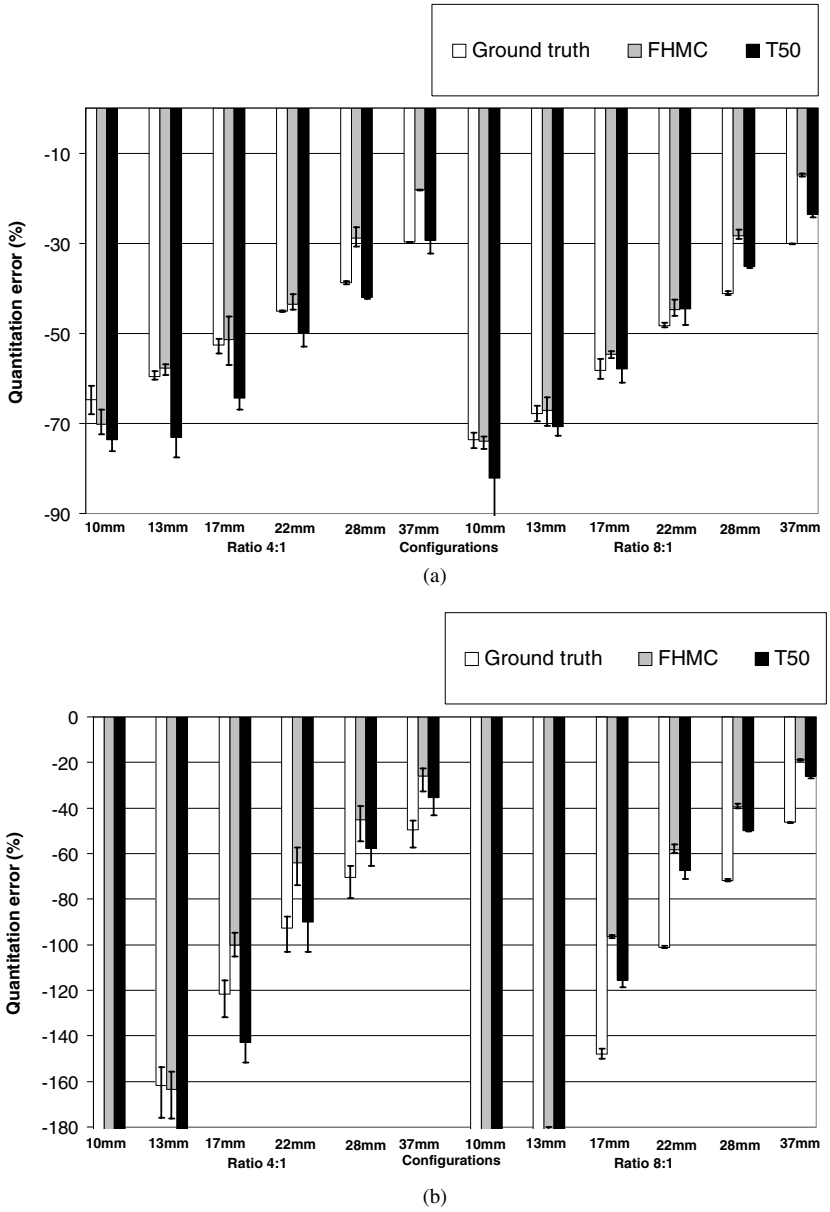



**Figure 9.** Lesion average activity concentration estimation errors from the simulated images: (a) 64 $mm^3$ voxels and (b) 8 $mm^3$ voxels for FHMC 0/2 versus T50 segmentation under variable imaging conditions.

reconstructed voxel sizes. As can be seen from this figure, FHMC and T50 led to the best results in comparison to the 'ground truth' throughout the different lesion sizes and contrasts evaluated, although T50 introduces larger errors in comparison to the 'ground truth' for lesion sizes of $<22$ mm and a contrast of 4:1. The use of the 8 $mm^3$ voxels does not alter the conclusions as far as the relationship between the results for the two methods evaluated is concerned, although in absolute terms all algorithms perform worse in comparison to the results obtained for 64 $mm^3$ voxels.

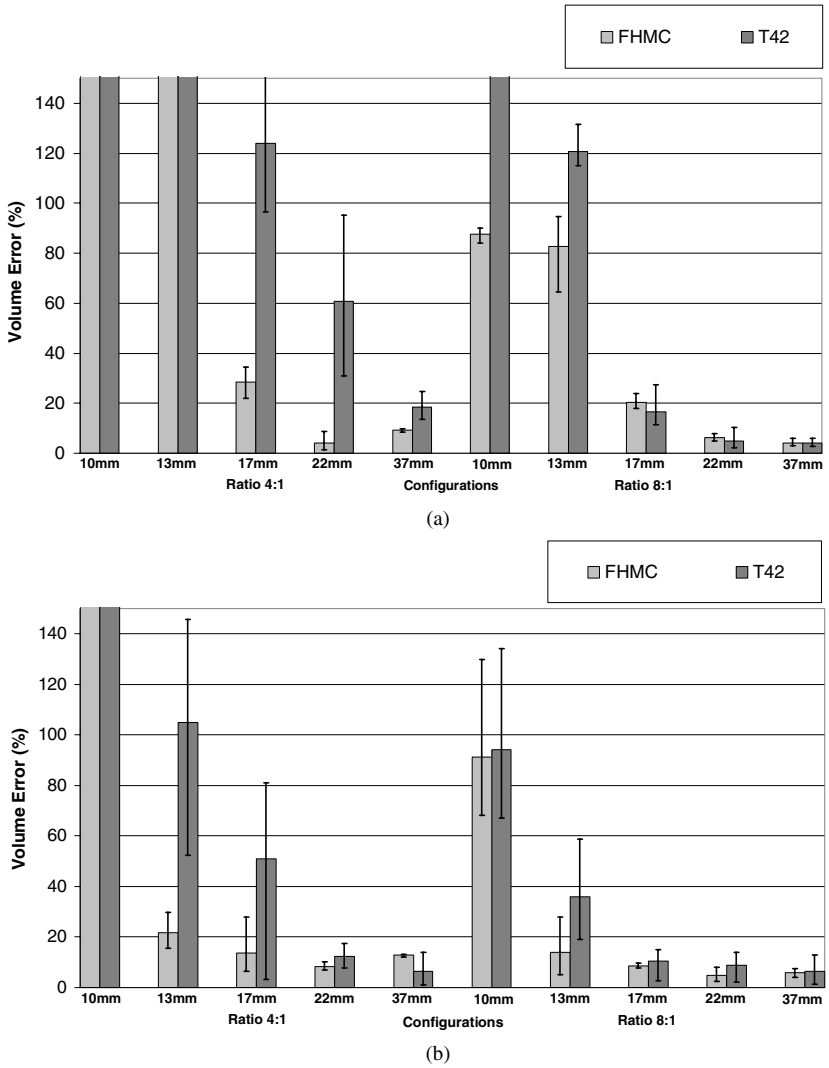



**Figure 10.** Lesion VOI estimation errors from the acquired images: (a) 64 $mm^3$ voxels and (b) 8 $mm^3$ voxels for the FHMC 1/2 versus T42 segmentation under variable imaging conditions.

Considering the acquired datasets, figures 10(a) and (b) contain the results for the % overall lesion volume estimation for the 64 $mm^3$ and 8 $mm^3$ voxels, while figures 11(a) and (b) show the corresponding results for the activity quantitation errors. In terms of the volume estimation the general trends were similar to those observed for the simulated datasets, with the FHMC performing better than the T42 throughout the range of lesion sizes and contrasts evaluated. In absolute terms, the FHMC results were better particularly in the case of 8 $mm^3$ voxels where errors of $<20\%$ and 10% were seen for lesions $>10$ mm and $>22$ mm respectively. T42 errors were similar to FHMC for the 8:1 ratio and spheres $>13$ mm but ranged from 20 to $>100\%$ for the 4:1 ratio configuration. A larger dependence on the statistical quality of the reconstructed images can be observed with the acquired datasets, demonstrating the more robust performance of the FHMC algorithm in comparison to the T42 methodology which was seen to be more affected by the images' statistical quality. Using again the example of the

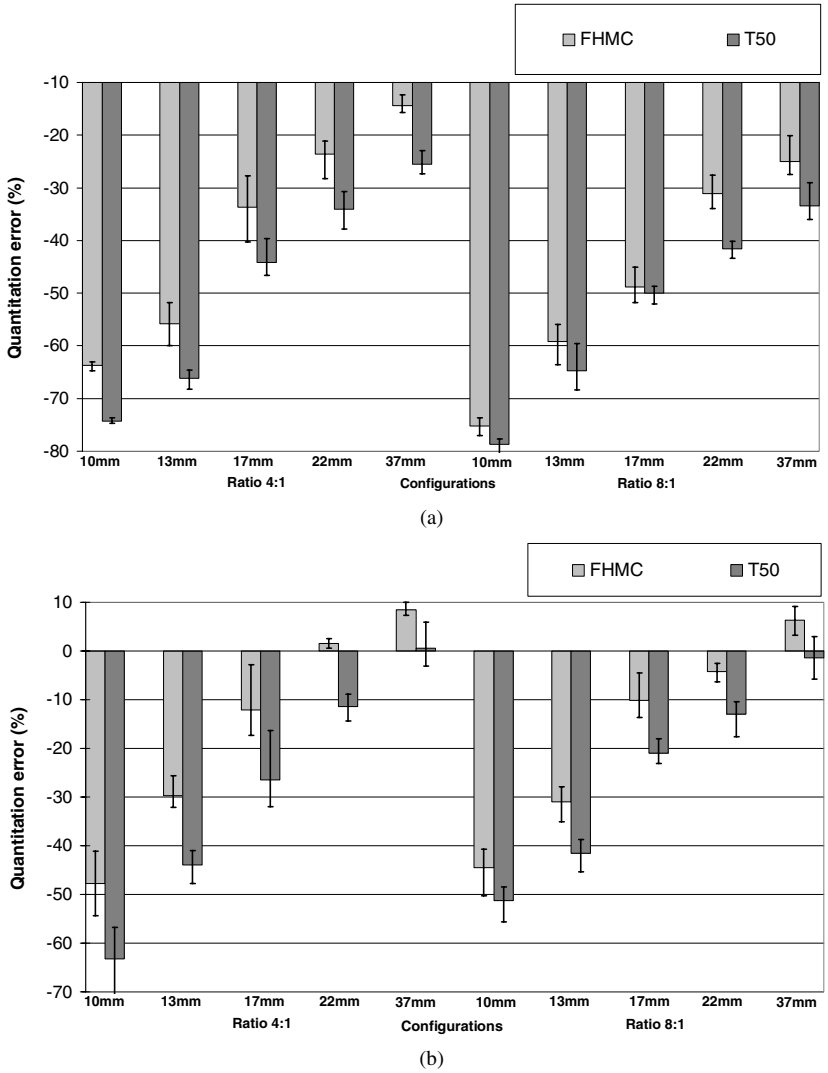



**Figure 11.** Lesion average activity concentration estimation errors from the acquired images: (a) 64 $mm^3$ voxels and (b) 8 $mm^3$ voxels for the FHMC 0/2 versus T50 segmentation under variable imaging conditions.

22 mm sphere (figure 10(a)), T42 errors were from 30 to 95% while FHMC errors were less than 5%. Although the variation of the FHMC results was higher for smaller spheres (10 and 13 mm), it was still smaller than in the case of the T42 results. For example, FHMC applied to the 13 mm sphere with a 4:1 contrast ratio (figure 10(b)) resulted in errors between 5 and 30% whereas T42 errors ranged from 50 to 150%. Similar results between the FHMC and the T50 algorithms were globally seen in terms of the % accuracy of the recovered activity concentration, confirming the trends observed with the simulated datasets. Finally, similarly with the volume estimation, better results were seen with the 8 $mm^3$ reconstructed voxel's size for both the T50 and the FHMC leading to activity concentration estimation errors of between +10% and −10% for lesions >17 mm in diameter.

## 4. Discussion

Although PET imaging applications are currently, in their majority, diagnostic and largely based on visual interpretation, there is increasing interest in applications such as the use of PET for radiotherapy treatment planning, as well as response to therapy and outcome prediction studies where accurate functional volume and concentration of activity estimation respectively are indispensable. Current state-of-the-art methodologies for functional volume determination involve the use of adaptive thresholding based on anatomical information or phantom studies. The performance of these techniques is greatly dependent on lesion contrast and image noise characteristics and as this work has demonstrated can lead to variable performance. On the other hand, already proposed automatic segmentation methodologies have been mostly evaluated for use in lesion detection rather than lesion volume determination. In addition, their performance is highly dependent, similarly to the thresholding algorithms, on image contrast and noise characteristics.

Hidden Markov chains is an automatic segmentation algorithm that allows noise modelling in the images but has also previously been evaluated for lesion detection rather than functional volume estimation. In the presented work a new algorithm (Fuzzy HMC) has been introduced and evaluated allowing the incorporation within hidden Markov chains of a finite number of fuzzy levels in combination with the 'hard' classes considered in HMC, adding this way an estimation of imprecision that can account for the effects of limited spatial resolution in emission tomography images. During the evaluation of the FHMC, the inclusion of more than two fuzzy levels was found to not substantially alter the segmentation results, while only the inclusion of the voxels belonging to the first fuzzy level led to the most accurate results in terms of functional volume calculations throughout the range of configurations considered. Although it would be possible to consider the use of HMC with four hard classes and an additional rule to cluster the resulting segmentation map, the fuzzy nature of the borders leads to computation issues in transition probabilities that HMC is not able to deal with. Note that the significant addition of the fuzzy measure and mathematical changes in the model allows FHMC to take into account such a configuration, mainly due to the fact that one given voxel can contain both classes.

Finally, in this paper we have introduced the concept of classification errors rather than volume estimation errors in the evaluation of segmentation algorithms for volume determination tasks. An evaluation based on classification errors is more robust since it does not simply consider the segmented volume but also its location with respect to the 'ground truth' known in simulated datasets. Therefore, while the absolute segmented volume may be correct its location may be wrong, a fact that is as significant as the correct estimation of the overall functional volume particularly for applications such as the use of PET volumes in radiotherapy treatment planning.

In comparison to the recommended T42 for the accurate functional volume determination in PET (Krak *et al* 2005), the FHMC algorithm gave clearly superior results for lesions $<28$ mm, in particular considering a lesion contrast of 4:1, where the T42 methodology completely failed to recover the functional volume. In addition, FHMC was more robust considering the different image statistical quality levels evaluated, while the results of T42 were greatly influenced by the level of noise present in the images. Differences between classification and volume estimation errors across the different configurations evaluated were smaller for the segmented volumes provided by the FHMC algorithm. In addition, the classification error results allow us to establish that the accuracy obtained on the estimated volumes using the FHMC algorithm is not by chance due to a similar level of negative and positive classification errors. A smaller reconstructed voxel size at the same statistical quality

level led to worse overall segmentation results, without altering the conclusions as far as the relative performance of the different algorithms evaluated is concerned.

The performance of the segmentation algorithms under evaluation in the reconstructed images of the acquired datasets was similar to that obtained with the simulated datasets in terms of volume estimation errors. The only difference observed in comparison with the simulated dataset results was an improvement in the overall segmentation results for 8 $mm^3$ reconstructed images in comparison to the 64 $mm^3$, which can be attributed to an associated adjustment of the optimized reconstruction parameters as a function of the voxel size in the 3D RAMLA algorithm used to reconstruct the acquired datasets.

As far as concentration recovery results are concerned, the performance of the different segmentation algorithms was compared in the simulated datasets to the recovered activity concentration considering the exact size and location of the simulated lesions. These results were influenced by the effects of partial volume leading to increasing underestimation of the activity concentration with decreasing lesion size. Segmentation algorithms concentrate on accurate edge modelling in the object of interest and do not as such account for changes in the values of the voxels as a result of PVE. FHMC and the current 'state-of-the-art' threshold of 50% of the maximum lesion value (Krak *et al* 2005) led to similar results independently of the configurations evaluated, with absolute differences of 10–15% (due to an extra underestimation for the T50 results). Similar trends to those observed with the simulated datasets were obtained from the segmentation of the acquired images.

The presented results demonstrate the interest of FHMC over thresholding algorithms as the flexibility of the fuzzy levels choice may allow the use of the same segmentation map for different tasks, across a large range of lesion contrasts and sizes. FHMC through the addition of the fuzzy levels associated with each hard class is able to more accurately model the object of interest edges in reconstructed PET images. In addition, FHMC is clearly less susceptible to alterations in statistical image quality and lesion contrasts than other methodologies. This was observed on both images of simulated and acquired datasets. Having said that, none of the evaluated algorithms was successful in accurate volume estimation for lesion sizes of $<17$ mm, considering typical PET image statistical qualities and reconstructed voxels of either 8 $mm^3$ or 64 $mm^3$. The main reason behind the failure of FHMC concerning the segmentation of such small lesions is the small number of voxels associated with the object of interest in combination with image noise levels, and the Hilbert–Peano path used to transform the image into a chain. The spatial correlation of such small objects may be lost once the image is transformed into a chain. A local model may be able to overcome such an issue (Hatt *et al* 2007).

The results for FHMC may be further improved. Firstly, the direct estimation of the noise in the reconstructed images may lead to better results in comparison to the assumed Gaussian model used in this work to fit the distribution for each of the classes. Secondly, other *a priori* models may be used for Markovian modelling, like couple (Pieczynski and Derode 2004) or triplet (Lanchantin and Pieczynski 2004) Markov chains or fields. These may be of interest considering a better modelling of the transitions between boundary classes, as well as the non-stationary nature of the hidden *a priori* model. In addition, the fuzzy model may be extended to more than two hard classes to better model inhomogeneous or non-spherical objects of interest.

## 5. Conclusion

A modified version of the hard Markov chains segmentation algorithm has been developed by introducing a fuzzy measure (FHMC). Our results with both simulated and acquired datasets

have shown that FHMC is more effective than the reference thresholding methodologies for both VOI determination and quantification in PET imaging. As part of the evaluation process, we have also introduced and assessed the interest of classification errors as a robust measurement of the performance of segmentation algorithms for VOI determination in contrast to a simple volume estimation which may introduce biases in terms of the segmented lesion location. Future developments will concentrate on the use of more than two 'hard' classes in FHMC, which may more accurately account for the presence of inhomogeneous or non-spherical functional volumes, as well as an investigation into more adequate noise and *a priori* models.

**Acknowledgment**

This work is financially supported by a Region of Brittany research grant under the 'Renouvelement des competences' programme 1202-2004.

## Appendix. The FHMC algorithm step by step

For the calculation of the expressions a quantization of the interval $[0, 1]$ into intervals $\left\{0, \frac{1}{N}, \frac{2}{N}, \ldots, \frac{N-1}{N}, 1\right\}$ is used. For example with two fuzzy levels (or intervals) $F_1, F_2$, we have $N = 3$ and there are $N - 1 = 2$ fuzzy levels with $\varepsilon_i = \frac{i}{N}$: $\varepsilon_1 = \frac{1}{3}$ and $\varepsilon_2 = \frac{2}{3}$. Note that the symbol $\widetilde{\ldots}$ denotes a density instead of a probability.

**(1) Transformation of the 2D or 3D image into a 1D chain using the Hilbert–Peano path (Kamata *et al* 1999) (save the path to be used in step 5 of the procedure).**

From this point on, every step is performed on the image transformed into a chain.

**(2) Parameters initialization**

*A priori* model parameters:

$\left.\begin{matrix}\alpha_1 = \alpha_2 = 0.25 \\ \gamma_1 = \gamma_2 = 0\end{matrix}\right\}$ $\text{init}(c)$ and $\text{trans}(c, d)$ are computed according to (3), (4) and the following:

$$\lambda = \frac{2}{N}\left(\sum_{i=1}^{N-1}\left(1 - \frac{i}{N}\right) + \sum_{i=1}^{N-1}\left(1 - \left|\frac{i}{N} - 1\right|\right)\right) + \frac{1}{N^2}\sum_{i=1}^{N-1}\sum_{j=1}^{N-1}\left(1 - \left|\frac{i}{N} - \frac{j}{N}\right|\right)$$

$$\beta = \frac{1 - (\alpha_1 + \alpha_2 + \gamma_1 + \gamma_2)}{\lambda}.$$

Initial and transition probabilities initializations can then be computed as follows:

$$\text{init}(0) = \alpha_1 + \gamma_1 + \frac{\beta}{N}\sum_{i=1}^{N-1}\left(1 - \frac{i}{N}\right)$$

$$\text{init}(1) = \alpha_2 + \gamma_2 + \frac{\beta}{N}\sum_{i=1}^{N-1}\left(1 - \left|1 - \frac{i}{N}\right|\right)$$

$$\text{init}(\varepsilon_i) = \frac{\beta}{N}\left((1 - \varepsilon_i) + (1 - |\varepsilon_i - 1|) + \frac{1}{N}\sum_{j=1}^{N-1}\left(1 - \left|\varepsilon_i - \frac{j}{N}\right|\right)\right)$$

$$\text{trans}(c, d) = C\frac{g(c, d)}{h(d)} \qquad \text{with} \quad \begin{cases} C = 1 & \text{if} \quad d \in \{0, 1\} \\ C = \frac{1}{N} & \text{if} \quad d \in ]0, 1] \end{cases}$$

*Noise model parameters:*

$(\{\mu_0, \mu_1\}, \{\sigma_0, \sigma_1\}) = K\,means(Y, 2)$ with $Y$ the image and 2 for the two hard classes to look for. Then we determine parameters of each fuzzy level with (9).

**(3) SEM procedure for parameters estimation**

At each iteration $q$ until no significant modification of the estimated parameters (convergence):

(a) *$\widetilde{fwd}$ and $\widetilde{bwd}$ densities computation for each class $c$, $c \in \{0, 1, \varepsilon_i\}$, $i = 1, \ldots, N-1$ is performed recursively as follows:*

$$\text{for } t = 1 : \widetilde{fwd}_1(c) = h(c)G_c(y_1)$$

$$\text{for } t > 1 : \widetilde{fwd}_t(c) = G_c(y_t)\left(\widetilde{fwd}_{t-1}(0)\widetilde{\text{trans}}\,(c, 0) + \widetilde{fwd}_{t-1}(1)\widetilde{\text{trans}}\,(c, 1) + \frac{1}{N}\sum_{i=1}^{N-1}\widetilde{fwd}_{t-1}(\varepsilon_i)\widetilde{\text{trans}}\,(c, \varepsilon_i)\right)$$

$$\text{for } t = T : \widetilde{bwd}_T(c) = 1$$

$$\text{for } t < T : \widetilde{bwd}_t(c) = G_0(y_{t+1})\widetilde{bwd}_{t+1}(0)\widetilde{\text{trans}}\,(0, c) + G_1(y_{t+1})\widetilde{bwd}_{t+1}(1)\widetilde{\text{trans}}\,(1, c) + \frac{1}{N}\sum_{i=1}^{N-1} G_{\varepsilon_i}(y_{t+1})\widetilde{\text{trans}}\,(\varepsilon_i, c)\widetilde{bwd}_{t+1}(\varepsilon_i).$$

These computations must be normalized. $G_c$ is given by: $G_c(y) = \frac{1}{\sigma_c\sqrt{2\pi}}\exp\left(-\frac{(y-\mu_c)^2}{2\sigma_c^2}\right)$.

(b) *Stochastic re-estimation of parameters:*

To obtain one *a posteriori* realization of X, simulate a fuzzy Markov chain using the following:

Posterior distributions of X are defined by:

$$\widetilde{\text{init}}(c) = \widetilde{fwdbwd}_1(c) \quad \text{and} \quad \widetilde{\text{trans}}\,(c, d)^{t+1} = \frac{f(d|c)G_d(y_{t+1})\widetilde{bwd}_{t+1}(d)}{\int_0^1 f(d|c)G_d(y_{t+1})\widetilde{bwd}_{t+1}(d)dv(d)}$$

$$\text{init}(c) = \begin{cases} \widetilde{\text{init}}(\text{c}) \text{ if } c \in \{0, 1\} \\ \frac{1}{N}\widetilde{\text{init}}\,(\text{c}) \text{ if } c \in ]0, 1] \end{cases} \quad \text{and} \quad \text{trans}(c, d)^t = \begin{cases} \widetilde{\text{trans}}\,(\text{c, d})^{\text{t}} \text{ if } d \in \{0, 1\} \\ \frac{1}{N}\widetilde{\text{trans}}\,(\text{c, d}) \text{ if } d \in ]0, 1]. \end{cases}$$

It has to be noted that $\text{trans}(c, d)^t$ depends on $t$ since a different transition matrix is computed for each element of the posterior realization, as we are dealing with a non-stationary Markov chain.

The estimated values of the parameters at the iteration $q$ are computed on the simulated *a posteriori* chain $\{x_t | t = 1, \ldots, T\}$ as follows:

$$\text{init}(c)^{[q]} = \delta\left(x_1^{[q]}, c\right).$$

For the *a priori* model: $$\text{trans}(c, d)^{[q]} = \frac{\sum_{t=2}^{T}\delta\left(x_{t-1}^{[q]}, c\right)\delta\left(x_t^{[q]}, d\right)}{\sum_{t=2}^{T}\delta\left(x_{t-1}^{[q]}, c\right)}.$$

For the noise model: $$\mu_c^{[q]} = \frac{\sum_{t=1}^{T} y_t\delta\left(x_t^{[q]}, c\right)}{\sum_{t=1}^{T}\delta\left(x_t^{[q]}, c\right)} \quad \sigma_c^{2[q]} = \frac{\sum_{t=1}^{T}\delta\left(x_t^{[q]}, c\right)\left(y_t - \mu_c^{[q]}\right)^2}{\sum_{t=1}^{T}\delta\left(x_t^{[q]}, c\right)}$$

for $c = 0$ and $c = 1$. For fuzzy levels ($c = \varepsilon_i$) noise parameters, use equation (9)

$$\text{with} \quad \delta(m, n) = \begin{cases} 1 & \text{if} \quad m = n \\ 0 & \text{if} \quad m \neq n. \end{cases}$$

**(4) MPM segmentation of the chain using estimated parameters**

For each $x_t$, determine the class (hard class or fuzzy level) minimizing the error classification probability by minimizing the following expression:

$$\widetilde{fwdbwd}_t(0)L(0,\hat{s}(y_t)) + \widetilde{fwdbwd}_t(1)L(1,\hat{s}(y_t)) + \int_0^1 \widetilde{fwdbwd}_t(\varepsilon_i)L(\varepsilon_i,\hat{s}(y_t))\,\mathrm{d}\varepsilon_i$$

for every $\hat{s}(y_t)$, and where $\widetilde{fwdbwd}$ denotes the product of the forward and backward densities. The cost function $L$ is given by (8).

**(5) Reverse transformation of the 1D segmented chain into the 2D or 3D segmentation map using the path saved at step 1.**

## References


Benmiloud B and Pieczynski W 1995 Estimation des paramètres dans les chaînes de Markov cachées et segmentation d'images *Traitement du signal* **12** 433–54

Celeux G and Diebolt J 1986 L'algorithme SEM: un algorithme d'apprentissage probabiliste pour la reconnaissance de mélanges de densités *Revue de Statistique Appliquée* **34** 35–52

Chen J L, Gunn S R and Nixon M S 2001 Markov random field model for segmentation of PET images *Lecture Notes Comput. Sci.* **2082** 468–74

Dai J 1994 Hybrid approach to speech recognition using hidden Markov models and Markov chains *Proc. Inst. Electron. Eng: Vis. Image Signal Process.* **141** 273–9

Delignon Y, Marzouki A and Pieczynski W 1997 Estimation of generalized mixtures and its application in image segmentation *IEEE Trans. Image Process.* **6** 1364–75

Dempster A P, Laird N M and Rubin D B 1977 Maximum likelihood from incomplete data via the EM algorithm *J. R. Stat. Soc.* B **39** 1–38

Erdi Y E, Mawlawi O, Larson S M, Imbriaco M, Yeung H, Finn R and Humm J L 1997 Segmentation of lung lesion volume by adaptative positron emission tomography image thresholding *Cancer* **80** 2505–9

Hatt M, Roux C and Visvikis D 2007 3D fuzzy adaptive unsupervised Bayesian segmentation for volume determination in PET *4th IEEE Int. Symp. on Biomedical Imaging: from Nano to Macro (Arlington, USA)*

International Electrotechnical Commission (IEC) 1998 *IEC Publication 61675-1: Radionuclide Imaging Devices—Characteristics and Test Conditions: Part 1. Position Emission Tomographs*

Jarritt H, Carson K, Hounsel A R and Visvikis D 2006 The role of PET/CT scanning in radiotherapy planning *Br. J. Radiol.* **79** S27–35

Kamata S, Eason R O and Bandou Y 1999 A new algorithm for N-dimensional Hilbert scanning *IEEE Trans. Image Process.* **8** 964–73

Kim J, Feng D D, Cai T W and Eberl S 2002 Automatic 3D temporal kinetics segmentation of dynamic emission tomography image using adaptative region growing cluster analysis *IEEE NSS-MIC Conf. Rec.* vol 3 pp 1580–3

Krak N C *et al* 2005 Effects of ROI definition and reconstruction method on quantitative outcome and applicability in a response monitoring trial *Eur. J. Nucl. Med. Mol. Imaging* **32** 294–301

Lamare F, Turzo A, Bizais Y, Cheze Le Rest C and Visvikis D 2006 Validation of a Monte Carlo simulation of the Philips Allegro/GEMINI PET systems using GATE *Phys. Med. Biol.* **51** 943–62

Lanchantin P and Pieczynski W 2004 Unsupervised non stationary image segmentation using triplet Markov chains *Advanced Concepts for Intelligent Vision Systems (ACVIS 04)* (Belgium)

Marroquin J, Mitter S and Poggio T 1987 Probabilistic solution of ill-posed problems in computational vision *J. Am. Stat. Assoc.* **82** 76–89

McQueen J 1967 Some methods for classification and analysis of multivariate observations *Proc. the 5th Berkeley Symp. on Mathematical Statistics and Probability* vol 1 pp 281–97

Nestle U, Kremp S, Schaefer-Schuler A, Sebastian-Welch C, Hellwig D, Rübe C and Kirsch C M 2005 Comparison of different methods for delineation of 18F-FDG PET-positive tissue for target volume definition in radiotherapy of patients with non-small cell lung cancer *J. Nucl. Med.* **46** 1342–8

Pieczynski W 2003 Modèles de Markov en traitement d'images *Traitement du Signal* **20** 255–77

Pieczynski W and Derode S 2004 Signal and image segmentation using pairwise Markov chains *IEEE Trans. Signal Process.* **52** 2477–89

Reader A J *et al* 2002 One-pass list-mode EM algorithm for high-resolution 3D PET image reconstruction into large array *IEEE Trans. Nucl. Sci.* **49** 693–9

Reutter B W, Klein G J and Huesman R H 1997 Automated 3D segmentation of respiratory-gated PET transmission images *IEEE Trans. Nucl. Sci.* **44** 2473–6

Riddell C, Brigger P, Carson R E and Bacharach S L 1999 The watershed algorithm: a method to segment noisy PET transmission images *IEEE Trans. Nucl. Sci.* **46** 713–19

Salzenstein F, Collet C and Petremand M 2004 Champs de Markov flous pour images multispectrales *Traitement du Signal* **21** 37–54

Salzenstein F and Pieczynski W 1997 Parameter estimation in hidden fuzzy Markov random fields and image segmentation *Graph. Models Image Process.* **59** 205–20

Salzenstein F and Pieczynski W 1998 Sur le choix de méthode de segmentation statistique d'images *Traitement du Signal* **15** 119–28

Visvikis D, Turzo A, Gouret S, Damine P, Lamare F, Bizais Y, Cheze Le and Rest C 2004 Characterisation of SUV accuracy in FDG PET using 3D RAMLA and the Philips Allegro PET scanner *J. Nucl. Med.* **45** 103P

Visvikis D *et al* 2001 Influence of OSEM and segmented attenuation correction in the calculation of standardised uptake values for FDG PET *Eur. J. Nucl. Med. Mol. Imaging* **28** 1326–35

Zhu W and Jiang T 2003 Automation segmentation of PET image for brain tumors *IEEE Nuclear Science Symp. Conf. Rec.—NSS-MIC* vol 4 pp 2627–9

# A Fuzzy Locally Adaptive Bayesian Segmentation Approach for Volume Determination in PET

Mathieu Hatt*, *Member, IEEE*, Catherine Cheze le Rest, Alexandre Turzo, Christian Roux, *Fellow, IEEE*, and Dimitris Visvikis, *Senior Member, IEEE*

***Abstract*—Accurate volume estimation in positron emission tomography (PET) is crucial for different oncology applications. The objective of our study was to develop a new fuzzy locally adaptive Bayesian (FLAB) segmentation for automatic lesion volume delineation. FLAB was compared with a threshold approach as well as the previously proposed fuzzy hidden Markov chains (FHMC) and the fuzzy C-Means (FCM) algorithms. The performance of the algorithms was assessed on acquired datasets of the IEC phantom, covering a range of spherical lesion sizes (10–37 mm), contrast ratios (4:1 and 8:1), noise levels (1, 2, and 5 min acquisitions), and voxel sizes (8 and 64 $mm^3$). In addition, the performance of the FLAB model was assessed on realistic nonuniform and nonspherical volumes simulated from patient lesions. Results show that FLAB performs better than the other methodologies, particularly for smaller objects. The volume error was 5%–15% for the different sphere sizes (down to 13 mm), contrast and image qualities considered, with a high reproducibility (variation $<4\%$). By comparison, the thresholding results were greatly dependent on image contrast and noise, whereas FCM results were less dependent on noise but consistently failed to segment lesions $<2$ cm. In addition, FLAB performed consistently better for lesions $<2$ cm in comparison to the FHMC algorithm. Finally the FLAB model provided errors less than 10% for nonspherical lesions with inhomogeneous activity distributions. Future developments will concentrate on an extension of FLAB in order to allow the segmentation of separate activity distribution regions within the same functional volume as well as a robustness study with respect to different scanners and reconstruction algorithms.**

***Index Terms*—Oncology, positron emission tomography (PET), segmentation, volume determination.**

Manuscript received June 03, 2008; revised December 02, 2008. First published January 13, 2009; current version published May 28, 2009. This work was supported in part by a Region of Brittany research grant under the "Renouvellement des compétences" under program 1202-2004, in part by the French National Research Agency (ANR) under Contract ANR-06-CIS6-004-03, and in part by the Cancéropôle Grand Ouest under the Contract R05014NG. *Asterisk indicates corresponding author.*

*M. Hatt is with the LaTIM, INSERM, U650, 29609 Brest, France and also with the Telecom Institute, Telecom Bretagne, 29609 Brest, France (e-mail: hatt@univ-brest.fr).

C. Cheze le Rest and A. Turzo are with the LaTIM, INSERM, U650, 29609 Brest, France and also with the Nuclear Medicine Department of Morvan Hospital, 29609 Brest, France (e-mail: catherine.cheze-lerest@chu-brest.fr; alexandre.turzo@chu-brest.fr).

C. Roux are with the LaTIM, INSERM, U650, 29609 Brest, France and also with the Telecom Institute, Telecom Bretagne, 29609 Brest, France (e-mail: christian.roux@telecom-bretagne.eu).

D. Visvikis is with the LaTIM, INSERM, U650, 29609 Brest, France (e-mail: Visvikis.Dimitris@univ-brest.fr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMI.2008.2012036

## I. Introduction

POSITRON emission tomography (PET) is now a widely used tool in the field of oncology, especially in applications such as diagnosis, and more recently radiotherapy planning [1] or response to therapy and patient follow-up studies [2]. On the one hand, accurate activity concentration recovery is crucial for correct diagnosis and monitoring response to therapy. On the other hand, applications such as intensity-modulated radiation therapy (IMRT) treatment planning using PET also require accurate shape and volume determination of the lesions of interest, in order to reduce collateral damage to healthy tissues and to ensure maximum dose delivered to the active disease. Various methodologies used for the determination of volume of interest (VOI) have been proposed. On the one hand, segmentation methods requiring a manual delineation of the boundaries of the object of interest have been established as laborious and highly subjective [2]. Alternatively, the performance of already available automatic algorithms is hampered by the low resolution and associated partial volume effects (PVE), as well as low contrast and signal to noise ratios generally characterizing PET images.

Most of the previously proposed work dealing with VOIs determination in PET use thresholding, either adaptive, based on *a priori* computed tomography (CT) knowledge [3], or a fixed threshold using values derived from phantom studies (from 30%–75% of maximum local activity concentration value) [1]–[3]. Thresholding is however known to be significantly susceptible to noise and contrast variations, leading to variable VOIs determination as shown in recent clinical studies [4]. As far as automatic detection of lesions from PET datasets is concerned, different methodologies have been previously proposed including edge detection [5], watersheds [6], fuzzy C-Means [7], or clustering [8]. The performance of these algorithms is also sensitive to variations in lesion-to-background contrast and/or noise levels. In addition, past work has in its majority considered the ability of such automatic methodologies for the detection of lesions (sensitivity), and not for their performance in terms of accuracy in the specific VOI determination task. Finally, all of the aforementioned algorithms have additional drawbacks associated with necessary preprocessing or postprocessing steps. For example in the case of the watershed algorithm, a preprocessing step using a filtering pass is required to smooth the image, and a postprocessing step is necessary to fuse the regions resulting from the over-segmentation of the algorithm. Such a need for user-dependent initializations, preprocessing and postprocessing steps, or additional information

like CT or expert knowledge render the use of these algorithms more complicated and the outcome dependent on choices made by the user in relation to these necessary steps.

Bayesian-based image segmentation methods are automatic algorithms allowing noise modelling and have shown to be less sensitive to noise than other segmentation approaches due to their statistical modelling [9]. They offer an unsupervised estimation of the parameters needed for the image segmentation and limit the user's input to the number of classes to be searched for in the image. Reconstructed images require no further preprocessing or postprocessing treatment (such as for example filtering) prior to the segmentation process. Instead, image noise is considered as additional information (a parameter in the classification decision process) to be taken into account rather than to be filtered or ignored. They have only been previously used in PET imaging in the form of hidden Markov fields (HMFs) [10] and more recently we have investigated the performance of hidden Markov chains (HMCs) for volume determination, a faster model that was in addition extended to include fuzzy modelling, fuzzy HMC (FHMC) [11]. Although FHMC was shown to provide overall superior results relative to the threshold reference methodology, independent of lesion contrast and image signal-to-noise ratio, it is unable to correctly segment objects $<2$ cm in diameter. This is mainly due to the 3-D Hilbert-Peano path [12] used to transform the 3-D volume into an 1-D chain, since voxels defining small objects may find themselves far away from each other on the chain, thus being misidentified by the algorithm as noise and becoming not significant enough to form a class apart from the background.

Consequently, the main objectives of this study were to improve the segmentation of small objects by 1) developing a fuzzy local adaptive Bayesian (FLAB) model and 2) comparing the performance of this new algorithm with that of the thresholding methodologies currently used in clinical practice as well as the fuzzy C-Means (FCM) and the previously proposed FHMC algorithms. In addition, as a secondary objective we have also investigated the use of the Pearson's system [13] in order to potentially improve the noise modelling used in the algorithm, instead of simply assuming a Gaussian distribution.

Different imaging conditions were considered in this study in terms of statistical quality, as well as lesion size and source-to-background (S/B) ratio. The images were reconstructed using an iterative algorithm, since this type of reconstruction algorithms form today's state of the art in whole body PET imaging in routine clinical oncology practice [14]. In addition, the new FLAB algorithm was evaluated using simulated images of non homogeneous and non spherical tumors derived from tumors of patients undergoing radiotherapy.

## II. Material and Methods

### A. FLAB Model

The FLAB model is an unsupervised statistical methodology that takes place in the Bayesian framework. Let $T$ be a finite set corresponding to the voxels of a 3-D PET image. We consider two random processes $\mathrm{Y} = (y_t)_{t\in T}$ and $\mathrm{X} = (x_t)_{t\in T}$. Y represents the observed image and takes its values in $\mathbb{R}$ whereas X represents the "hidden" segmentation map and takes its values in the set $\{1,\ldots,C\}$, with C being the number of classes. The segmentation problem consists of estimating the hidden X from the available noisy observation Y. The relationship between X and Y can be modeled by the joint distribution $P(\mathrm{X},\mathrm{Y})$, which can be obtained using the Bayes formula

$$P(\mathrm{X}|\mathrm{Y}) = \frac{P(\mathrm{X},\mathrm{Y})}{P(\mathrm{Y})} = \frac{P(\mathrm{Y}|\mathrm{X})P(\mathrm{X})}{P(\mathrm{Y})}. \tag{1}$$

$P(\mathrm{Y}|\mathrm{X})$ is the likelihood of the observation Y conditionally with respect to the hidden ground-truth X, and $P(\mathrm{X})$ is the prior knowledge concerning X. The Bayes rule allows the determination of the posterior distribution of X with respect to the observation $\mathrm{Y}: P(\mathrm{X}|\mathrm{Y})$. Contrary to the FHMC model [11], we do not assume here that a Markov process can model the prior distribution of X, thus simplifying its expression.

*The Fuzzy Measure:* The general idea behind the implementation of a fuzzy model within the Bayesian framework was previously introduced in [15] and [16] and was used for a local Bayesian segmentation scheme in [15]. Its implementation is based on the incorporation of a finite number of fuzzy levels $F_i$ in combination with two homogeneous (or "hard") classes, in comparison to the standard implementation where only a finite number of hard classes are considered. This model allows the coexistence of voxels belonging to one of two hard classes and voxels belonging to a "fuzzy level" depending on its membership to the two hard classes. While the statistical part of the algorithm models the uncertainty of the classification, with the assumption being that the voxel is identified but the observed data is noisy, the fuzzy part models the imprecision of the voxel's membership, with the assumption being that the voxel may contain both classes. One way to achieve this extension is to simultaneously use Dirac and Lesbegue measures, considering that X in the fuzzy model takes its values in [0,1] instead of $\Omega = \{1,\ldots,C\}$. We define therefore a new measure $\nu = \delta_0 + \delta_1 + \zeta$ on [0,1], given that $\delta_0$ and $\delta_1$ are the Dirac measures at 0 and 1, and $\zeta$ is the Lesbegue measure on the fuzzy interval [0,1]. This approach is adapted for the segmentation of PET images since they are both noisy and of low resolution. The "noise" aspect when considering Bayesian models is the way the values of each class to be found in the image are distributed around a mean value. The noise model used, whose respective mean and variance are to be determined by the estimation steps, can therefore be adapted to image specific characteristics. Finally, the fuzzy measure facilitates a more realistic modelling of the objects' borders transitions between foreground and background, allowing in such a way to indirectly account for the effects of blurring associated with low resolution PET images.

*Distribution of X (a priori model):* Using $\nu = \delta_0 + \delta_1 + \zeta$ as a measure on [0,1], the *a priori* distribution of each $x_t$ can be defined by a density $h$ on [0,1], with respect to $\nu$. If we assume that X is a stationary process and that the distribution of each $x_t$ is uniform on the fuzzy class, this density can be written as

$$\begin{aligned} h(0) &= P[x_t = 0] = p_0 \\ h(1) &= P[x_t = 1] = p_1 \\ h(\varepsilon) &= P[x_t = \varepsilon] = 1 - p_0 - p_1 \text{ for } \varepsilon \in [0,1] \end{aligned} \tag{2}$$

where, $h$ satisfies the following normalization condition:

$$h(0) + h(1) + \int_0^1 h(\varepsilon)d\varepsilon = 1.$$

Using this simple modelling for the prior distribution leads to ignoring the spatial relationship of each voxel with respect to its local neighborhood. Although it is possible to include such spatial information using the contextual framework [15], the use of such modelling leads to an increase in the number of parameters to be handled, and in practice, no more than one or two neighbors can be actually taken into account. Hence, the contextual approach is not of interest since we aim to explore all the information available in the 3-D volume around each voxel, i.e., at least 26 neighbors (eight-connectivity extended in three dimensions). As an alternative, the adaptive framework [15] can be used. In this adaptive modelling, the spatial information is inserted into the estimation step of the algorithm (see section *parameters estimation*).

*Distribution of Y (Observation or Noise Model) and the Pearson's System:* In order to define the distribution of Y conditional on X, let us consider two independent random variables $\mathrm{Y}_0$ and $\mathrm{Y}_1$, associated with the two "hard" values 0 and 1, whose densities $f_0$ and $f_1$ are characterized by means and variances $(\mu_0, \sigma_0^2)$ and $(\mu_1, \sigma_1^2)$ respectively. The mean and variance of each fuzzy level $F_i$ are derived from the ones estimated in the two hard classes as follows:

$$\begin{aligned}\mu_{F_i} &= \mu_0(1-\varepsilon_i) + \varepsilon_i\mu_1 \\ \sigma_{F_i}^2 &= \sigma_0^2(1-\varepsilon_i)^2 + \varepsilon_i^2\sigma_1^2\end{aligned} \tag{3}$$

where $\varepsilon_i$ is the value associated to a fuzzy level $F_i$. For the case of two fuzzy levels $\varepsilon_1 = 1/3$ and $\varepsilon_2 = 2/3$ were used according to results previously published [11].

The assumption that the noise for each class of the observed data can fit a Gaussian distribution was considered as a first approximation as with the previous implementation of the FHMC algorithm [11]. In this work, we propose the study of the Pearson's system that contains seven other distributions. In this context, instead of using a Gaussian distribution, an additional step is introduced to detect which laws best fit the actual distribution of the voxels in the image, for each class considered at a given iteration of the estimation step of the algorithm. The theory behind the Pearson's system has been previously detailed in [17] and a description of its use in mixture estimation and statistical image segmentation is given in [13]. Here, we briefly describe the Pearson's system in our particular context.

A distribution density $f$ on $\mathbb{R}$ belongs to the Pearson's system if it satisfies

$$\frac{1}{f(y)}\frac{df(y)}{dy} = -\frac{y+a}{c_0 + c_1 y + c_2 y^2}. \tag{4}$$

Different shapes of distributions as well as the parameters determining a given distribution are provided by the variations of the coefficients $a$, $c_0$, $c_1$, and $c_2$. For $m = 1$, 2, 3, and 4, let us consider the first four statistical moments of a partition $\mathrm{Y_p}$ of Y defined by

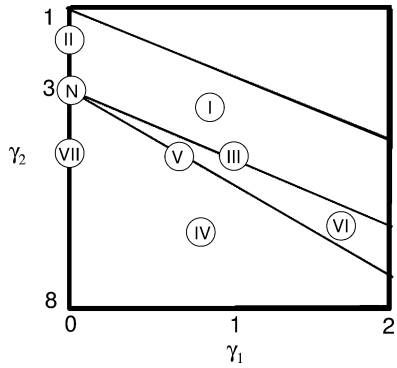



Fig. 1. The eight distribution families in the graph of Pearson, function of $\gamma_1$ and $\gamma_2$ [17]. I for Beta I, II for type II, III for Gamma, IV for type IV, V for Inverse Gamma, VI for Beta II, VII for type VII and N for Normal.

$$\begin{aligned}\mu_1 &= E[\mathrm{Y_p}] \\ \mu_p &= E\left[(\mathrm{Y_p} - E[\mathrm{Y_p}])^m\right] \text{ for } \mathrm{m} \geq 2.\end{aligned} \tag{5}$$

We also define two parameters $\gamma_1$ and $\gamma_2$ as follows:

$$\gamma_1 = \frac{\mu_3^2}{\mu_2^3} \text{ and } \gamma_2 = \frac{\mu_4}{\mu_2^2} \tag{6}$$

where $\sqrt{\gamma_1}$ is called "skewness" and $\gamma_2$ is called "kurtosis." The coefficients $a$, $c_0$, $c_1$, and $c_2$ are related to (5) and (6) by equations that can be found in the Appendix I-A. Given $\lambda = \gamma_1(\gamma_2+3)^2/4(4\gamma_2 - 3\gamma_1)(2\gamma_2 - 3\gamma_1)(2\gamma_2 - 3\gamma_1 - 6)$, the eight distribution density families $\{f_1, \ldots, f_8\}$ contained in the system of Pearson can now be defined by a set of conditions using $\lambda$, $\gamma_1$ and $\gamma_2$ (see Appendix I-B). These eight distribution density families are illustrated in Fig. 1. Finally, the protocol used for the determination of which density family best fits each measured distribution can be found in Appendix I-C.

*Parameters Estimation:* The different parameters necessary to be estimated for the segmentation process are

$$\begin{aligned}\omega &= (A, B) \\ A &= (p_0, p_1) \\ B &= \left(\mu_0, \sigma_0^2, \mu_1, \sigma_1^2\right).\end{aligned} \tag{7}$$

Both *a priori* $(A)$ and noise $(B)$ parameters are unknown and may vary from one image to another. An iterative procedure called stochastic expectation maximization (SEM) [18], a stochastic version of the EM algorithm [19], is used for the estimation of these parameters. This is achieved by sampling a realization of X according to its posterior distribution $P(\mathrm{X}|\mathrm{Y})$ and computing empirical values of the parameters of interest using this realization. The stochastic nature of this procedure makes it less sensitive to the initial guess of the parameters using the K-Means [20] than deterministic procedures like the EM algorithm. The system of Pearson can be used as an additional step (inside each iteration of the algorithm) in order to determine the type of distribution to use. The posterior distribution $d$ with respect to class $c$ for a given voxel $t$ used at iteration $q$ for sampling the posterior realization is given by (8) at the bottom of the

page, where, $f^{q-1}(y_t|c)$ is a density whose distribution type is chosen using the Pearson system and whose mean and variance were estimated at iteration $q-1$, and $p_{t,c}^{q-1}$ is the prior probability of voxel $t$ belonging to class $c$ estimated at iteration $q-1$.

In the adaptive framework priors are reestimated using a local neighboring window with priors $p_{t,c}$ depending on the position $t$ of the voxel in the image. Although in the 2-D case, a window centred on the voxel of interest is used [15], for our application we use a 3-D "cube" centred on each voxel. The size of the estimation "cube" was experimentally determined for the specific application of PET imaging, since it depends on the size of the objects of interest (10–50 mm in diameter) relative to the reconstructed voxel size ($2\times2\times2$ or $4\times4\times4$ mm$^3$). An estimation cube should from one hand be small enough to yield good local characteristics [15], while on the other hand it should not be too large with respect to the size of the object of interest. Considering this, we tested two different estimation "cube" sizes; namely covering $3\times3\times3$ and $5\times5\times5$ voxels.

It is worth noting that only the priors are concerned by the use of the adaptive framework. Noise parameters are estimated the same way as in the blind context [15]. The detailed description of the SEM algorithm in our context is given in the Appendix II.

*Segmentation:* In order to perform segmentation on a voxel by voxel basis, we need to use a criterion to classify each voxel as either part of the background or the functional VOI. For this purpose we use the maximum posterior likelihood (MPL) method as suggested by [15]. To compute a solution, the MPL method requires the parameters defining the *a priori* model (priors of each class for each voxel) as well as the noisy observation data model (mean and variance of each class), estimated using SEM. The MPL computes the posterior density and selects for each voxel the class that maximizes it, using the procedure described below.

Let us consider $d(\varepsilon|y_t)$ given by (8) computed using the parameters estimated by the SEM estimation algorithm. Using $d(F|y_t)=1-d(0|y_t)-d(1|y_t)$, the decision rule assigning the class $c$ or fuzzy level $F_i$ to the voxel $t$ knowing the observed value $y_t$ is given by the following procedure:

For each voxel, let $c_t=\arg\max_{n\in\{0,1,F\}} d(n|y_t)$. If $c_t\in\{0,1\}$, then assign the hard class 0 ($c_t=0$) or 1 ($c_t=1$) to the voxel $t$. Else if $c_t$ belongs to the fuzzy domain ($c_t=F$), use $c_t=\arg\max_{n\in]0,1[} d(n|y_t)$ to determine its exact value using the quantitation of the fuzzy interval into fuzzy levels (see Section II-A-3) and assign one of the fuzzy levels to the voxel. In our implementation of FLAB, each $c_t$ can take four different values: 0, 1/3, 2/3, and 1.

### B. Alternative Approaches Used for Comparison

*Thresholding:* Various thresholding methodologies have been proposed in the past for functional volume determination [2]–[4]. For comparison purposes with the developed methodology, threshold at 42% of the maximum value inside the lesion was chosen for VOI determination, based on suggestions from previous publications [2], [3]. The methodology was implemented through region growing using the voxel of maximum intensity in the object of interest as a seed. Using a 3-D neighborhood (26 neighbors) the region is iteratively increased by adding neighboring voxels if their intensity is superior or equal to the selected threshold value. The results derived using this method will be denoted from here onwards as T42.

*Fuzzy C-Means:* The FCMs algorithm was introduced in [21]. It was suggested for PET image segmentation in [7]. For the purpose of this study it was implemented using the following objective function O:

$$O(\varepsilon)=\sum_{i=1}^{I}\sum_{j=1}^{J}\varepsilon_{ij}^{e}|\varepsilon_{ij}-m_i|^2 \tag{9}$$

where $e\geq1$ is a weighting exponent and $m_i$ are the centre values of the classes. The weighting exponent $e$ controls the fuzzy aspect of the image and is usually set to 2 (hard segmentation is represented by $e=1$). The algorithm converges to the value at which the objective function has a local maximum. The results derived using this method will be denoted from here onwards as FCM.

### C. Validation Studies

*Datasets:* Acquisitions of the IEC image quality phantom [22], containing six different spherical lesions of 10, 13, 17, 22, 28, and 37 mm in diameter [Fig. 3(a)] were carried out in list-mode format using a Philips GEMINI PET/CT scanner. The spatial resolution of this system is 4.9 mm full-width at half-maximum (FWHM) at the center of the field of view [23]. Partial volume effects are therefore expected to be significant even for the largest sphere. The 28-mm-diameter sphere was not considered in this study since it was replaced by a handmade plastic sphere whose diameter was not known precisely. Different parameters were considered covering a large spectrum of configurations allowing assessment of the influence of different parameters susceptible to affect the functional VOI determination. The statistical quality of the images was varied by considering 1, 2, or 5 min list-mode time frames. Two different signal-to-background (S/B) ratios (4:1 and 8:1) were considered, by introducing 7.4 kBq/cm$^3$ in the background and 29.6 or 59.2 kBq/cm$^3$, respectively, in the spheres. Two different voxel sizes ($2\times2\times2$ or $4\times4\times4$ mm$^3$) were used in the reconstruction of each of the different statistical quality datasets using the 3-D RAMLA algorithm, with specific parameters previously optimized for clinical use [14]. Visual illustration of the acquired images is given in Fig. 2. In addition, an estimation of

$$d^q(c|y_t)=\frac{p_{t,c}^{q-1}f^{q-1}(y_t|c)}{p_{t,0}^{q-1}f^{q-1}(y_t|0)+p_{t,1}^{q-1}f^{q-1}(y_t|1)+\left(1-p_{t,0}^{q-1}-p_{t,1}^{q-1}\right)\int_0^1 f^{q-1}(y_t|\theta)d\theta} \tag{8}$$

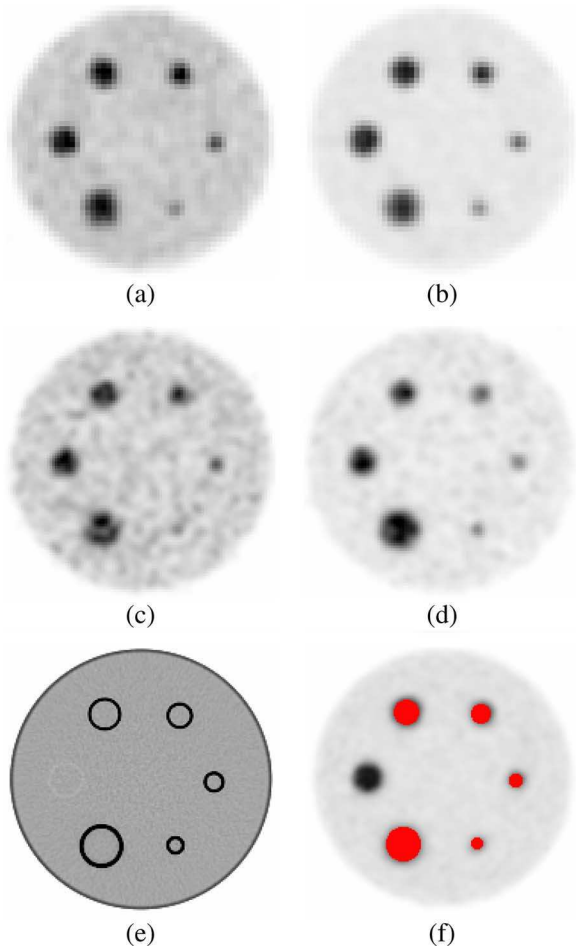



Fig. 2. Different images used in the segmentation study; (a) ratio 4:1, 2 min acquisition time, 64 $mm^3$ voxels, (b) ratio 8:1, 2 min, 64 $mm^3$, (c) ratio 4:1, 2 min, 8 $mm^3$, (d) ratio 8:1, 2 min, 8 $mm^3$, (e) CT acquisition, (f) voxel-by-voxel ground-truth generated using CT image on the PET image. Note the 28 mm sphere is in plastic and not clearly seen (since its real diameter was unknown this sphere was excluded from any analysis in this work).

the FLAB algorithm's reproducibility was performed by considering five different 1 min list-mode time frames acquired consecutively and reconstructed using 8 $mm^3$ voxels.

Finally, to test the algorithm against more clinically realistic conditions of tumor shapes, we simulated three lesions with nonspherical shapes and inhomogeneous activity distributions. These lesions were generated using real lung tumor images from three patients undergoing $^{18}$FDG PET scans for radiotherapy treatment planning purposes. A ground-truth was drawn by a nuclear medicine physician (on a slice-by-slice basis) based on the reconstructed patient images. In the case of the first tumor, the simulated contrast between the region of the highest activity concentration and the rest of the tumor was around 2.2:1 whereas in the case of the second tumor, it is closer to 1.4:1. Finally, the third tumor is almost homogeneous. The overall contrast between the whole tumor and the background was 6:1 and 5:1 for the first and second tumors, respectively, and less than 2:1 for the third one. In terms of lesion size, the largest lesion "diameter" was 4.1, 2.9, and 1.5 cm for the first, second, and third lesion, respectively. These lesions were subsequently placed within the lungs of the NCAT phantom [24]. No respiratory or cardiac motion was considered. Normal organ FDG concentration was assumed for the simulation [25], with the maximum activity concentration in the lesions being four times the mean activity concentration in the lungs. The NCAT emission and attenuation maps were finally combined with a model of the Philips PET/CT scanner previously validated with GATE [26]. A total of 45 million coincidences were simulated corresponding to the statistics of a standard clinical acquisition over a single axial field of view of 18 cm [26]. Images were subsequently reconstructed from the list mode output of the simulation using 8 $mm^3$ voxels. As well as using all of the simulated true coincidences, images were reconstructed for each lesion using only 40% and 20% of the overall detected coincidences in order to evaluate the accuracy of the segmentation algorithms at different noise levels (similar to the IEC phantom study using 5, 2, and 1 min acquisitions for the image reconstruction). Visual illustration of these simulated tumor images (central slice), with their ground-truth drawn from the corresponding patient tumors are displayed in Figs. 7 and 8, and Fig. 9(a)–(c). Each segmentation algorithm considered was applied to the lesion and the segmentation map was compared with the ground-truth. Note that in this framework, the ground-truth does not need to be accurate with respect to the true patient image. What is important is that we are able to compare the segmentation obtained on the simulated image with the ground-truth used in the simulation. The corresponding segmentation maps (central slice) for each algorithm can be found in Figs. 7 and 8 and Fig. 9(d)–(g).

*Analysis:* As our goal is not lesion detection in the whole body image but the estimation of a lesion's volume with the best accuracy possible, we assume that the lesion has been previously identified by the clinician and automatically or manually placed in a 3-D "box" well encompassing the object [see Fig. 3(a)]. Although no significant impact on the segmentation results was observed through small changes in placement or size of the box, certain conditions must be respected. Evidently it should be large enough to contain the entire extent of the object of interest and a significant number of background voxels so the algorithm is able to detect and estimate the parameters of the background class. On the other hand it should be small enough in order to avoid including neighboring tissues with significant uptake that would end up being classified as functional VOI, requiring manual postprocessing. However, the shape of this box does not have to be perfectly cubic or with specified dimensions (contrary to the FHMC case [11]), and as a result it could be drawn accordingly to exclude structures in the background that are of no interest.

Subsequently, the images of the selected area were segmented in two classes (functional VOI and background) using each of the methods under evaluation (T42, FCM, FHMC, and FLAB). In the FHMC and FLAB cases, considering the optimization results obtained in [11], two fuzzy levels were considered in the segmentation process and the functional volumes were defined using the first hard class and the first fuzzy level. A voxel-to-voxel ground-truth was generated for the phantom dataset using the CT image registered with the PET reconstructed image [see

Fig. 2(e) and (f)]. Classification errors (CE) were then computed on a voxel-by-voxel basis following the definition used in [11]:

$$\mathrm{CE} = \frac{(\mathrm{PCE} + \mathrm{NCE})}{\mathrm{VoS}} \times 100. \tag{10}$$

PCE stands for positive classification errors, including voxels of the background that are classified as belonging to the object of interest, and NCE stands for negative classification errors including voxels of the object that are classified as belonging to the background. These classification errors essentially occur on the boundaries of the objects of interest because of activity "spill in" and "spill out." If the segmentation results in PCEs and NCEs of equal amounts, the computed VOI would be very close to the true known volume whereas the shape and position of the object would be incorrect (this essentially occurs for objects $>2$ cm, while for smaller objects the errors are essentially PCE). As shown in (10), the total number of PCEs and NCEs is considered with respect to the number of voxels defining the sphere (VoS). Although the size of classification errors can be bigger than 100%, in the case where a large number of background voxels in the selected area of interest are misclassified as belonging to the sphere, maximum classification errors considered in this paper where limited to 100%, since any such values represent complete failure of the segmentation process. Although the combination of PCE and NCE into CE leads to a loss of information as far as the direction of the bias is concerned, classification errors represent more pertinent information than overall volume errors, which reflect neither accurate magnitude nor direction of the bias for a segmented volume. For comparison purposes overall volume errors (with respect to the known volume of the sphere) were also computed and shown in Fig. 6.

As far as the simulated tumors are concerned, both overall volume errors (with respect to the known volume of the ground-truth) and CE were computed. Since all the algorithms under investigation in this study perform binary segmentations (i.e., able to distinguish between tumor tissue and background only), no evaluation was performed of their ability to distinguish different regions within a given tumor.

## III. Results

Different segmentation maps obtained using each of the methods under evaluation (FHMC, FLAB, T42, and FCM) are presented in Fig. 3(c)–(f) for a slice centered on the 22 mm sphere considering a "good quality" image (8:1 contrast and 5 min acquisition) [Fig. 3(a)] to visually illustrate the variations of the segmentation maps obtained. Segmentation results in the case of a "lower quality" image (4:1 contrast and 2 min acquisition) and a smaller sphere (17 mm) [Fig. 3(g)] are presented in Fig. 3(h)–(k). Both images are representative of the 8 $mm^3$ voxel size reconstructions.

In the different figures shown in this section the CE are given for all five spheres (10, 13, 17, 22, and 37 mm) and for both contrast ratios (4:1 on the left part of each figure, 8:1 on the right part) considered. The error bars in the figures represent the different results obtained for each of the three different levels of image statistical quality considered. The top of the error bar is the result concerning the worst statistical quality images (1 min acquisition), the medium one concerns the medium quality (2 min acquisition), and the lowest one corresponds to the superior statistical quality (5 min acquisition). The only exception is Fig. 5 where the error bars represent the variability of the FLAB segmentation results considering the application of the algorithm on multiple images of 1 minute acquisitions (five independent realizations).

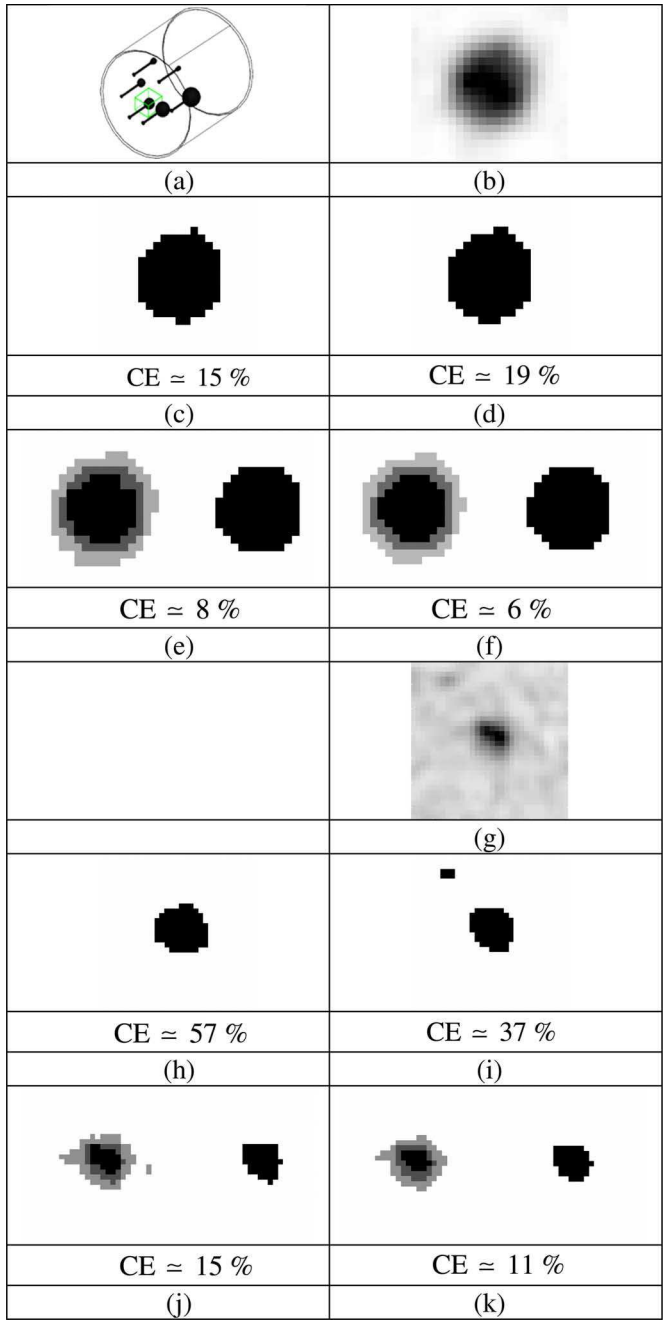



Fig. 3. (a) Graphical representation of the IEC phantom and illustration of the 3-D box selection for the 22-mm sphere and examples of segmentation maps (only central slice is shown); (b)–(f) for the 22 mm sphere (8:1 contrast, 5 min acquisition) and (g)–(k) for the 17 mm sphere (4:1 contrast, 2 min acquisition) with corresponding volume errors (computed on the whole volume): (b) and (g) PET ROI, (c) and (h) T42 map, (d) and (i) FCM map, (e) and (j) FHMC, and (f) and (k) FLAB maps with two fuzzy levels (light and dark grey voxels). Both images are extracted from 8 $mm^3$ voxel size reconstructions.

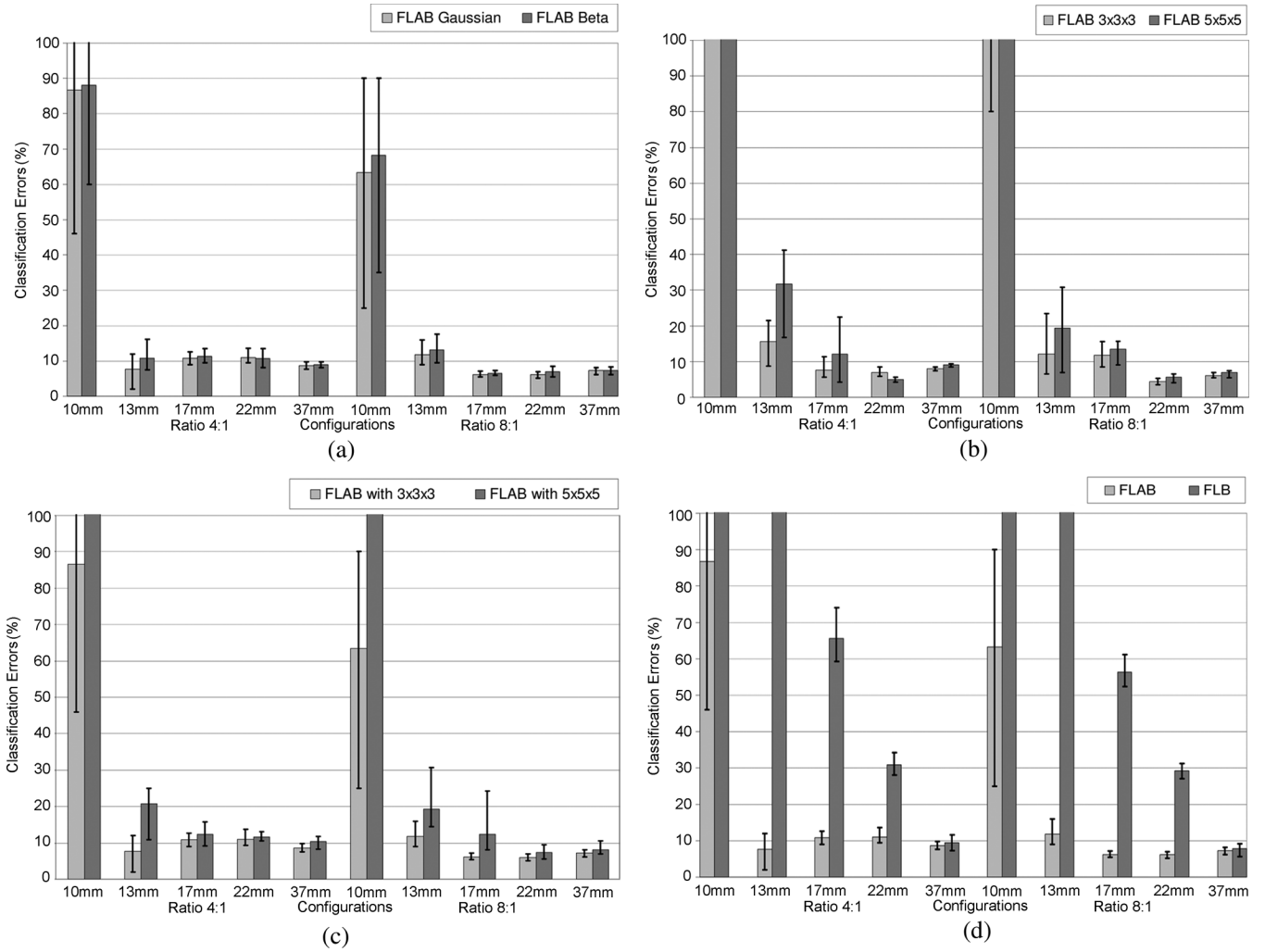



Fig. 4. Optimization of the FLAB algorithm. Classification errors for (a) Beta I distributions (detected using the Pearson's system) or Gaussian distributions (for the 8 $mm^3$ voxel size); (b) $3 \times 3 \times 3$ or $5 \times 5 \times 5$ voxels for the estimation cube (for the 64 $mm^3$ voxel size); (c) $3 \times 3 \times 3$ or $5 \times 5 \times 5$ voxels for the estimation cube (for the 8 $mm^3$ voxel size); (d) with (FLAB) or without (FLB) adaptive estimation of priors (for the 8 $mm^3$ voxel size). The top of the error bar is the result concerning the worst statistical quality images (1 min acquisition), the medium one concerns the medium quality (2 min acquisition), and the lowest one corresponds to the superior statistical quality (5 min acquisition).

Fig. 4 contains the results on the optimization of the algorithm for the specific application of lesion segmentation in PET images. Considering the selected volume of interest around a lesion, the Pearson's system systematically led to the detection of Beta I distributions for both the background and the lesion activity distributions (although with different parameters). However, the parameters $\gamma_1$ and $\gamma_2$ (see (6)) placed the estimated distributions very close to the Gaussian one in the Pearson graph [as it can be seen in Fig. 1, the surface matching Beta I distribution (I) is in contact with the point defining the Normal distribution (N)]. Consequently only small changes in the volume estimation results were consistently obtained using the Beta I instead of a Gaussian distribution [Fig. 4(a)]. Considering these results the Gaussian distribution was kept in the final implementation of the algorithm for the description of both the background and lesion activity distributions.

In terms of the size of the estimation "cube" used for the re-estimation of the priors in the adaptive framework, a size of $3 \times 3 \times 3$ voxels led to consistently better results across different lesion and voxel sizes as well as S/B contrast and noise configurations as shown in Fig. 4(b) and (c). Finally, Fig. 4(d) demonstrates the impact in terms of the improved results through the use of the adaptive estimation, for the 8 $mm^3$ configuration. In this figure the FLAB segmentation results are compared to the results without adaptive estimation (FLB for Fuzzy Local Bayesian, using the same fuzzy levels implementation), where priors are the same for all the voxels of the image and are computed using the entire image instead of using only the local neighbourhood of each voxel. As is demonstrated by this figure, the inclusion of the adaptive estimation significantly improves the segmentation results throughout the different lesion sizes and contrast configurations considered.

Results in relation to the FLAB algorithm's reproducibility can be seen in Fig. 5. In this particular figure, error bars represent the variation of the segmentation results (mean and variance) using the five different images obtained from the consecutive 1 minute acquisitions. A variation of $<4\%$ in the segmented volumes was obtained from the application of the algorithm on the five different images for all spheres except from the 1 cm sphere which the algorithm consistently failed to correctly seg-

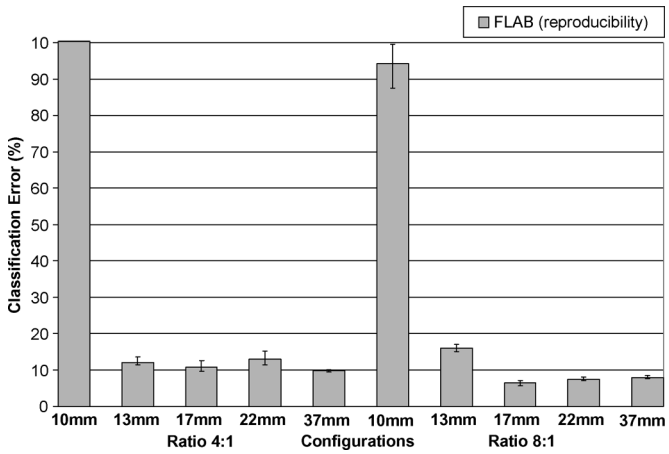



Fig. 5. Study of FLAB reproducibility using five different 1 min list-mode time frames (reconstructed with 8 $mm^3$ voxel size). The error bars represent the variability of the FLAB segmentation results considering the application of the algorithm on multiple images of 1 minute acquisitions (five realizations).

ment. This segmentation failure is most probably the cause of this larger variability observed for the segmented volume of the 1 cm sphere.

Fig. 6 presents the classification errors and corresponding overall volume errors relative to the CT-based ground-truth obtained using each approach, for both 64 and 8 $mm^3$ voxel sizes [Fig. 6(a) and (b) and 6(c) and (d), respectively]. Globally, volume errors are very closely linked to classification errors: when the segmentation results in strictly NCE, the volume error (underestimation) is equal to the CE. When the segmentation results in only PCE, the volume error (overestimation) is also equal to the CE. And when both NCE and PCE occur, the volume error is inferior to the CE (it essentially occurs for medium-sized spheres). FLAB led to superior results in comparison to all the other methodologies on the whole dataset. The proposed algorithm gives good results (on average between 5% and 20% CE) independently of the contrast ratio and for all spheres except from the 1 cm one for which a minimum error of 25% was obtained for the most favorable configuration evaluated (8:1 contrast and a 5 min. acquisition). The use of a reconstruction voxel size of $8mm^3$ allowed an improvement in the segmented volume errors from 10%–25% to 5%–15% for lesions between 1 and 2 cm.

As shown in Fig. 6, T42 gave errors $<20\%$ for the three biggest spheres with the 8:1 contrast and 64 $mm^3$ voxel size, while for a 4:1 contrast T42 did not manage to accurately segment any of the spheres. By reducing the reconstruction voxel size to 8 $mm^3$ an improvement was obtained in the results of the T42 with errors $<15\%$ for the three larger spheres and a contrast 8:1, while errors of $<20\%$ were obtained for the 22 and 37 mm spheres with a 4:1 contrast ratio. In the case of the FCM algorithm errors of $<20\%$ and $>40\%$ were seen for lesions larger and smaller than 2 cm, respectively. No substantial differences were seen in these results from the reduction in the reconstruction voxel sizes from 64 to 8 $mm^3$. Finally, FLAB performed better in comparison to the previously developed fuzzy Bayesian approach (FHMC) for all different lesion sizes and statistical image qualities considered with a larger magnitude effect (improvements of over 100% in the errors) observed in the spheres with a diameter $<2$ cm. Relative to the FLAB results globally larger improvements in the accuracy of the segmented volumes were observed for the FHMC algorithm with a reduction in the reconstructed voxel size. On the other hand, in percentage terms the dependence of the algorithm results to the statistical quality of the images was similar for both the FLAB and FHMC results.

Figs. 7–9 show visual illustration of the segmentation maps obtained on the simulated tumors. Fig. 10 contains the results for both classification errors (NCE+PCE divided by the number of voxels defining the tumor ground-truth volume) and volume errors (with respect to known overall volume of the tumor) for each approach.

The results for the first and third tumors (Fig. 7) show the largest differences between the four algorithms. In the case of the first tumor, this difference can be attributed to the nonuniform activity distribution (the contrast between the region of highest activity and the rest of the tumor is around 2.2:1) relative to the second tumor (closer to 1.4:1). Consequently, the segmentation results of T42 and FCM lead to large under evaluation ($-30$ to $-50\%$) of the true volume of the first tumor since they limit themselves to the highest activity area, whereas in the case of the second tumor they are unable to differentiate between the two regions, hence recovering the entire tumor (less than 10% error_for all methods). On the other hand, the third tumor despite being uniform is small with a low tumor to background ratio (1.5 cm in "diameter" and contrast $<2$:1). As a result, thresholding using 42% of maximum value fails completely (the region growing never stops and expands into the entire selection box) and FCM despite qualitatively satisfying results leads to a large over evaluation (from 10% to 40% volume error depending on the image statistical quality) of the volume. As far as FHMC and FLAB are concerned, they are both able to recover the whole tumor in all cases with volume errors between 2% and $<20\%$ (see Fig. 10). While FLAB in comparison with the FHMC performed better in terms of both the misclassification and the overall volume errors, FHMC results were less competitive with decreasing tumor sizes as seen also in the IEC phantom results (Fig. 10). Finally, the variability of the results (demonstrated by the error bars in Fig. 10) considering the different noise levels was higher for FCM and T42, illustrating their lower robustness to noise in comparison to the fuzzy statistical approaches.

## IV. Discussion

Over the past few years there has been an increasing interest in clinical applications such as the use of PET for IMRT planning, for which an accurate estimation of the functional volume is indispensable. Unfortunately, accurate manual delineation is impossible to achieve due to high inter- and intra-observer variability [2] resulting from the noisy and low resolution nature of the PET images. Current state of the art methodologies for functional volume determination involve the use of adaptive thresholding based on anatomical information or phantom studies. Thresholding however is known to be sensitive to contrast variation as well as noise [2], [4], since it does not include any explicit modelling of noise or spatial relationship.

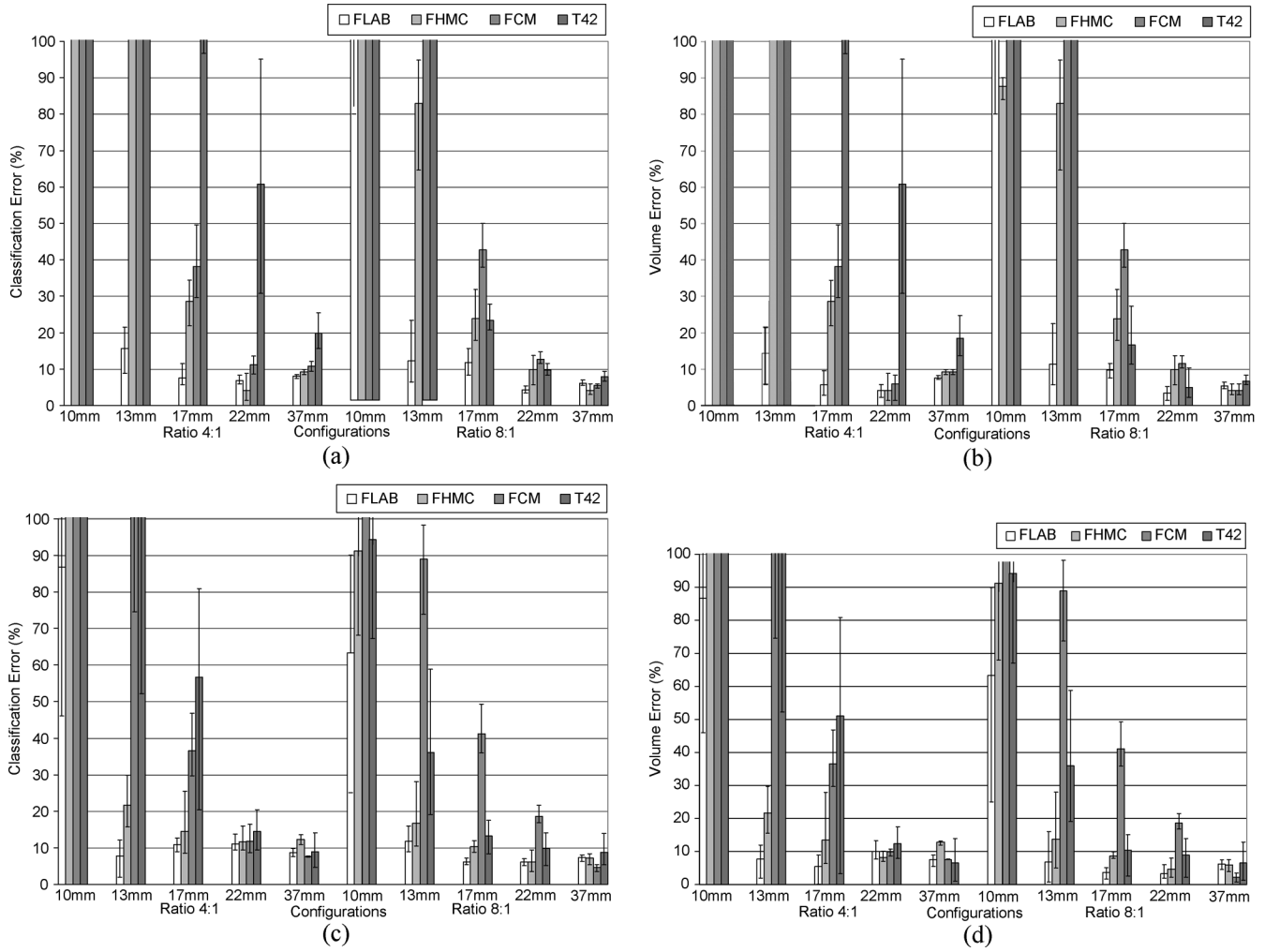



Fig. 6. Comparison of performances for FLAB, FHMC, FCM, and T42 on data reconstructed with (a) classification errors and (b) volume errors, for 64 $mm^3$ and (c) classification errors and (d) volume errors, for 8 $mm^3$ voxels. The top of the error bar is the result concerning the worst statistical quality images (1 min acquisition), the medium one concerns the medium quality (2 min acquisition), and the lowest one corresponds to the superior statistical quality (5 min acquisition).

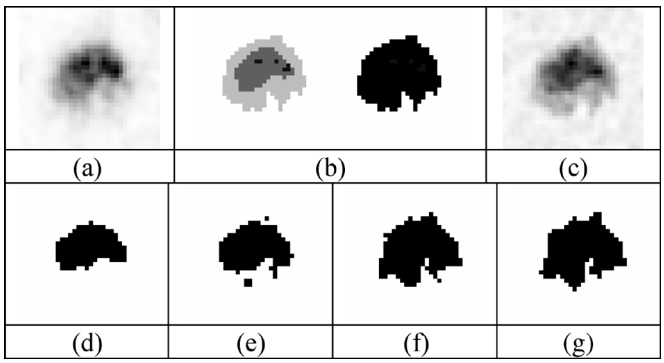



Fig. 7. (a) Real tumour used as model, (b) voxelized ground-truth (manually drawn) and its binary version, and (c) simulated tumour. Segmentation binary maps obtained using (d) T42, (e) FCM, (f) FHMC, and (g) FLAB are shown. Image is 34 $\times$ 34 voxels with 8 $mm^3$ voxels.

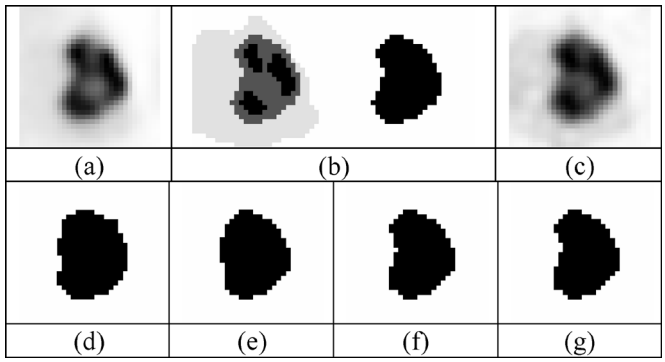



Fig. 8. (a) Real tumour used as model, (b) voxelized ground-truth (manually drawn) and its binary version, and (c) simulated tumour. Segmentation binary maps obtained using (d) T42, (e) FCM, (f) FHMC, and (g) FLAB are shown. Image is 30 $\times$ 30 voxels with 8 $mm^3$ voxels.

In addition, proposed adaptive thresholding methodologies require *a priori* knowledge of the tumor volumes currently obtained by CT images, based on the nonvalid assumption that the functional and anatomical volumes are the same [3]. In addition, proposed correction methodologies accounting for the effects of background activity levels depend on lesion contrast and background noise as well as being imaging system specific [4]. On the other hand, previously developed automatic algorithms have also shown dependence on the level of noise and lesion contrast, most frequently requiring preprocessing

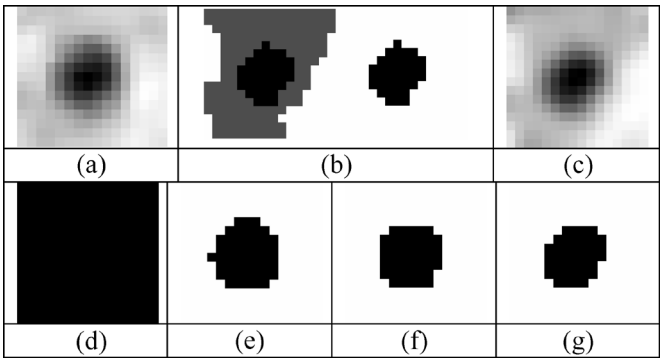



Fig. 9. (a) Real tumour used as model, (b) voxelized ground-truth (manually drawn) and its binary version, and (c) simulated tumour. Segmentation binary maps obtained using (d) T42, (e) FCM, (f) FHMC, and (g) FLAB are shown. Image is 16 × 16 voxels with 8 $\mathrm{mm}^3$ voxels.

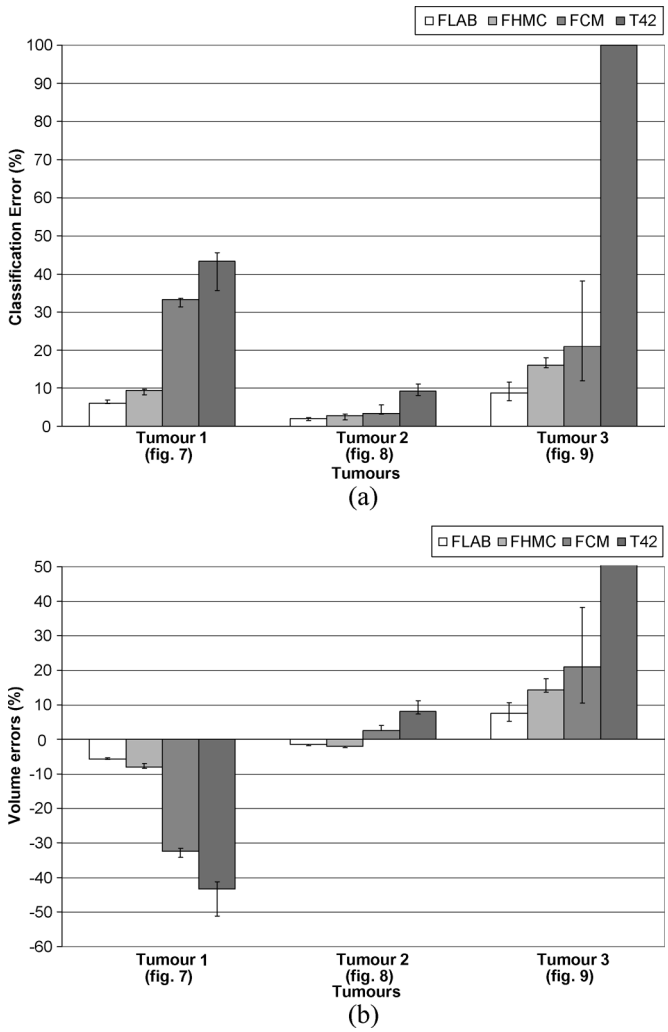



Fig. 10. Segmentation results for the three simulated tumours. (a) Classification errors and (b) overall volume errors. The top of the error bar is the result concerning the worst statistical quality images (20% of detected coincidences), the medium one concerns the medium quality (40%), and the lowest one corresponds to the superior statistical quality (100%).

or postprocessing steps and variable initialization parameter values depending on image characteristics rendering their use complicated and their performance highly variable.

We have previously developed and assessed the performance for functional volume segmentation of a modified version of the hidden Markov chains algorithm (FHMC) [11]. In this algorithm a number of fuzzy levels have been added to introduce the notion of imprecision allowing this way to account for the effects of low image spatial resolution in addition to the noise modelling (which is part of the standard HMC framework). Although the algorithm was shown to accurately segment functional volumes (errors $<15\%$) for lesions $>2$ cm throughout different contrast and noise conditions, it was unable to accurately segment lesions $<2$ cm. The main reason behind the failure of FHMC concerning the segmentation of such small lesions was the small number of voxels associated with the object of interest in combination to image noise levels, and the Hilbert-Peano path [12] used to transform the image into a chain. The spatial correlation of such small objects is lost once the image is transformed into a chain, because the voxels belonging to the object may find themselves far away from each other in the chain, thus resulting in transition probabilities that prevent these voxels to form a class differentiated from the background. In addition, it was thought that the assumption of a Gaussian noise distribution in the images to be segmented may have also been partly responsible.

FLAB clearly improved the results of FHMC, essentially due to the adaptive estimation of the priors using the whole 3-D neighborhood of each voxel, as the results of Fig. 5(c) clearly demonstrate. FLAB results obtained on the objects $>2$ cm were similar to those obtained through the use of FHMC as were their respective robustness with respect to noise levels. Finally, FLAB resulted in faster computation times in comparison with the FHMC.

In addition, highly reproducible results ($<4\%$ variability, to compare with the 8%–20% variability observed on manual segmentation [2]) were obtained for different image contrast ratios and lesion sizes $>1$ cm. We should emphasize here that the performance of the FLAB in comparison to other segmentation algorithms was evaluated in this study on images reconstructed using a specific iterative reconstruction algorithm used today in clinical practise. Since the FLAB segmentation algorithm has been developed in order to better cope with variable noise and contrast characteristics it should be the least affected by such changes introduced as a result of using an alternative reconstruction algorithm [27]. On the other hand, the use of the system of Pearson for the determination of image voxel value distributions did not lead to significant changes or improvements in the results in comparison to the Gaussian assumption. Although this was shown to be the case for the images reconstructed using the specific iterative reconstruction algorithm used here it may not be the case if an alternative reconstruction algorithm is used, where potentially the use of the system of Pearson for the characterisation of the image voxel values distribution may still prove to play a role in the segmentation process and needs to be further investigated.

By comparison the use of T42 led, as expected, to segmented functional volumes greatly dependent on image contrast and noise levels while being comparable to the FLAB results considering medium image statistical quality and lesions $>17$ mm with an 8:1 tumor to background ratio. Finally, the resulting vol-

umes from the application of the automatic segmentation algorithm FCM were less dependent to image statistical quality but consistently failed to segment lesions $<2$ cm.

In this study, as in every other phantom study presented to date in the literature, we have firstly considered the performance of the different algorithms for the segmentation of uniformly filled spherical lesions. To our knowledge there has been no study up to now specifically investigating the functional volume segmentation task for inhomogeneous uptake lesions, for example lesions with necrotic or partially necrotic regions. Although it has not been the major aim of their work, Nestle *et al.* demonstrated some evidence of the issues associated with the use of either fixed or background adjusted thresholding methodologies for lesions with inhomogeneous activity distributions and shapes in the clinical set up for non small cell lung cancer [4]. As it was shown in this study using simulated realistic lesions, the FLAB model is able to successfully deal with nonuniform lesion shapes and variable activity concentrations in contrast with the threshold based or fuzzy C-means segmentation algorithms considered. On the other hand, the binary two-class modelling (background or lesion) is obviously not adequate to permit the differentiation of multiple regions inside the tumor with largely different activity concentrations, as well as extracting the overall tumor in the case of strong heterogeneity. However, whereas it seems difficult to improve threshold-based segmentation methods in order to allow the identification of regions with variable activity concentration within the same functional volume of interest, the fuzzy model of FLAB may be extended to more than two hard classes to allow modelling a combination of inhomogeneous regions within a given volume. This could further enhance the use of FLAB for functional volume segmentation in future potential clinical applications.

The objectives of this study were to address the issue of functional volume determination and lesion segmentation. The FLAB model, as with any other segmentation algorithms, does not modify the values of the image voxels. As such, the use of the functional volume obtained with the FLAB algorithm, although is the closest to the true volume of the tumor as demonstrated by the results in this study, does not lead to the accurate activity concentration within the lesion. This is as a result of including voxels whose values have been decreased by spill-out from partial volume effects, usually leading to an underestimation of the activity concentration whose magnitude depends on the size of the lesion [11]. Although the segmented volume should therefore not be used for directly recovering the accurate activity concentration, they can be used in combination with partial volume correction methodologies potentially allowing a more accurate correction in comparison to the use of anatomical volumes [28].

## V. Conclusion

A modified version of a fuzzy local Bayesian segmentation algorithm has been developed. The suggested approach combines statistical and fuzzy modelling in order to address specific issues in the segmentation of low resolution noisy PET images. It is automatic, fully 3-D and uses adaptive estimation of priors to yield good local spatial characteristics that improve segmentation of small objects of interest. Results obtained with images of the IEC phantom reconstructed with the 3-D RAMLA iterative algorithm have shown that it is more effective than the reference thresholding methodology and other previously proposed automatic algorithms such as FHMC or the FCM methods for functional volume determination in PET images. The algorithm has also been tested successfully against realistic simulated tumors, using real patient tumors as model, with nonspherical shape and inhomogeneous activity distributions. Future developments will concentrate on the incorporation into FLAB of three hard classes and three different fuzzy transitions, in order to allow the segmentation within the same lesion of variable activity distributions in the case of highly heterogeneous functional uptake in the tumor volumes. We will also evaluate the use of different noise models in an associated robustness study using acquisitions with different scanner models and reconstruction algorithms.

## Appendix I

### A. Relationship Between Coefficients $a$, $c_0$, $c_1$, and $c_2$ and (5) and (6)

$$a=\frac{(\gamma_2+3)\sqrt{\gamma_1\mu_2}}{10\gamma_2-12\gamma_1-18}-\mu_1$$

$$c_0=\frac{\mu_2(4\gamma_2-3\gamma_1)-\mu_1(\gamma_2+3)\sqrt{\gamma_1\mu_2}+\mu_1^2(2\gamma_2-3\gamma_1-6)}{10\gamma_2-12\gamma_1-18}$$

$$c_1=\frac{(\gamma_2+3)\sqrt{\gamma_1\mu_2}-2\mu_1(2\gamma_2-3\gamma_1-6)}{10\gamma_2-12\gamma_1-18}$$

$$c_2=\frac{(2\gamma_2-3\gamma_1-6)}{10\gamma_2-12\gamma_1-18}$$

### B. Definition of the Eight Distribution Density Families

$$\begin{aligned}
&f\in f_1 \Leftrightarrow \lambda<0 && : \text{Beta of the first kind(I)}\\
&f\in f_2 \Leftrightarrow \gamma_1=0 \text{ and } \gamma_2<3 && : \text{Type II(II)}\\
&f\in f_3 \Leftrightarrow 2\gamma_2-3\gamma_1-6=0 && : \text{Gamma(III)}\\
&f\in f_4 \Leftrightarrow 0<\lambda<1 && : \text{Type IV(IV)}\\
&f\in f_5 \Leftrightarrow \lambda=1 && : \text{Inverse Gamma(V)}\\
&f\in f_6 \Leftrightarrow \lambda>1 && : \text{Beta II(VI)}\\
&f\in f_7 \Leftrightarrow \gamma_1=0 \text{ and } \gamma_2>3 && : \text{Type VII(VII)}\\
&f\in f_8 \Leftrightarrow \gamma_1=0 \text{ and } \gamma_2=3 && : \text{Normal(Gaussian)(N)}
\end{aligned}$$

Beta I and Gaussian distributions with respect to a class $c$ are defined as follows:

$$\text{Gaussian}_c(y)=\frac{1}{\sigma_c\sqrt{2\pi}}\exp\left(-\frac{(y-\mu_c)^2}{2\sigma_c^2}\right) \tag{18}$$

$$Beta_c(y)=\frac{1}{B(\alpha,\beta)}y^{\alpha-1}(1-y)^{\beta-1} \tag{19}$$

where $B(\alpha,\beta)=\Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha+\beta)$ is the Beta function (with $\Gamma$ the Gamma function).

We also have the following relationships between the parameters $\alpha$ and $\beta$, and the mean and variance ($\hat{\mu}_c$, $\hat{\sigma}_c^2$ denote estimated mean and variance) of class $c$ (this is useful to get the

parameters $\alpha$ and $\beta$ from the estimated means and variances obtained through the SEM algorithm)

$$\mu_c = \frac{\alpha}{\alpha+\beta}$$
$$\sigma_c^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$
$$\alpha = \hat{\mu}_c\left(\frac{\hat{\mu}_c(1-\hat{\mu}_c)}{\sigma_c^2}-1\right)$$
$$\beta = (1-\hat{\mu}_c)\left(\frac{\hat{\mu}_c(1-\hat{\mu}_c)}{\sigma_c^2}-1\right).$$

### C. Recipe for Identification of the Best Family to Fit Distributions of Classes

Let us consider the voxels $y_1,\ldots,y_t$ and their partitions $Q_0$ and $Q_1$ into two classes. The moments can be estimated from empirical moments, and we use the following to detect which family best fits each distribution.

1) Consider the partitions $Q_0, Q_1$ of $(x_1,\ldots,x_t)$ defined by $i \in Q_0 \Leftrightarrow x_i = 0$ and $i \in Q_1 \Leftrightarrow x_i = 1$.
2) For each class $i$ use $Q_i$ in order to estimate the $\mu_{m,i}$ empirical moments by the following.

$$\mu_{1,i} = \frac{\sum_{t\in Q_i} y_t}{\mathrm{Card}(Q_i)}$$
$$\mu_{p,i} = \frac{\sum_{t\in Q_i}(y_t-\mu_{1,i})^m}{\mathrm{Card}(Q_i)}$$

   for $m = 2, 3, 4$.
3) For each class $i$, calculate $\gamma_{1,i}$ and $\gamma_{2,i}$ from the estimated $\mu_{m,i}$ $(m = 1, 2, 3, 4)$ according to (6).
4) For each class $i$, use $\gamma_{1,i}$, $\gamma_{2,i}$ and rules (Appendix I-B) to determine which family its density $f$ belongs to.

## Appendix II

### SEM Algorithm

1) Give an initial value of the parameters

$$\omega^0 = [p_{t,0}^0, p_{t,1}^0, \mu_0^0, (\sigma_0^2)^0, \mu_1^0, (\sigma_1^2)^0]$$

   using K-Means algorithm for the noise and equal probabilities for the priors.
2) At each iteration $q$, $\omega^q$ is obtained from $\omega^{q-1}$ and the data $(y_1,\ldots,y_t)$ using:
   a) Choose a distribution for the classes 0 and 1 according to the Pearson system rules (Section II-A-3 and Appendixes I-B and I-C).

For each $y_t$, compute the *a posteriori* probabilities $d^q(0|y_t)$ and $d^q(1|y_t)$ using (8) and sample a value in the set $\{0, 1, F\}$ according to $d^q(0|y_t)$, $d^q(1|y_t)$ and $1-d^q(0|y_t)-d^q(1|y_t)$ ($F$ representing the fuzzy voxels). Let us denote $R = (r_1^q,\ldots,r_T^q)$ the posterior realization obtained through this sampling.

Let $Q_0^q = \{t|r_t^q = 0\}$ and $Q_1^q = \{t|r_t^q = 1\}$.

- Reestimate the priors using

$$p_{c,t}^q = \frac{1}{\mathrm{Card}(C_t)}\sum_{j\in C_t}\delta\left(r_j^q, \mathrm{c}\right) \text{ for } c \in \{0,1\}$$

  where $C_t$ is the estimation cube centred on voxel $t$ and $\delta(a,\mathrm{b})$ the Kronecker function.
- Reestimate the noise parameters using

$$\hat{\mu}_c^{q+1} = \frac{\sum_{t\in Q_c^q} y_t}{\mathrm{Card}\,(Q_c^q)}$$
$$(\hat{\sigma}_c^2)^{q+1} = \frac{\sum_{t\in Q_c^q}\left(y_t-\hat{\mu}_c^{q+1}\right)^2}{\mathrm{Card}\,(Q_c^q)} \text{ for } c \in \{0,1\}$$

For the means and variances of the fuzzy levels, use (3).

Repeat step 2 until stabilization of the parameters. Stabilization is defined by a criterion of % change in the values of the parameters between two successive iterations (we used 0.1% and the algorithm usually stops before 25 iterations) and a maximum number of iterations if the stabilization criterion is not met (usually 50 iterations).

## References


[1] H. Jarritt, K. Carson, A. R. Hounsel, and D. Visvikis, "The role of PET/CT scanning in radiotherapy planning," *Br. J. Radiol.*, vol. 79, pp. S27–S35, 2006.

[2] N. C. Krak and R. Boellaard *et al.*, "Effects of ROI definition and reconstruction method on quantitative outcome and applicability in a response monitoring trial," *Eur. J. Nucl. Med.Molecular Imag.*, vol. 32, pp. 294–301, 2005.

[3] Y. E. Erdi, O. Mawlawi, S. M. Larson, M. Imbriaco, H. Yeung, R. Finn, and J. L. Humm, "Segmentation of lung lesion volume by adaptative positron emission tomography image thresholding," *Cancer*, vol. 80, no. S12, pp. 2505–2509, 1997.

[4] U. Nestle, S. Kremp, A. Schaefer-Schuler, C. Sebastian-Welch, D. Hellwig, C. Rübe, and C. M. Kirsch, "Comparison of different methods for delineation of 18F-FDG PET-positive tissue for target volume definition in radiotherapy of patients with non-small cell lung cancer," *J. Nucl. Med.*, vol. 46, no. 8, pp. 1342–1348, 2005.

[5] B. W. Reutter, G. J. Klein, and R. H. Huesman, "Automated 3-D segmentation of respiratory-gated PET transmission images," *IEEE Trans. Nucl. Sci.*, vol. 44, no. 6, pp. 2473–2476, 1997.

[6] C. Riddell, P. Brigger, R. E. Carson, and S. L. Bacharach, "The watershed algorithm: A method to segment noisy PET transmission images," *IEEE Trans. Nucl. Sci.*, vol. 46, no. 3, p. 731, Jun. 1999, 719.

[7] W. Zhu and T. Jiang, "Automation segmentation of PET image for brain tumours," in *IEEE NSS-MIC Conf. Rec.*, Oct. 2003, vol. 4, pp. 2627–2629.

[8] J. Kim, D. D. Feng, T. W. Cai, and S. Eberl, "Automatic 3-D temporal kinetics segmentation of dynamic emission tomography image using adaptative region growing cluster analysis," in *IEEE NSS-MIC Conf. Rec.*, 2002, vol. 3, pp. 1580–1583.

[9] W. Pieczynski, "Modèles de Markov en traitement d'images," *Traitement du Signal*, vol. 20, no. 3, pp. 255–277, 2003.

[10] J. L. Chen, S. R. Gunn, and M. S. Nixon, "Markov random field model for segmentation of PET images," *Lecture Notes on Computer Science*, vol. 2082, pp. 468–474, 2001.

[11] M. Hatt, F. Lamare, N. Boussion, A. Turzo, C. Collet, F. Salzenstein, C. Roux, K. Carson, P. Jarritt, C. Cheze-Le Rest, and D. Visvikis, "Fuzzy hidden Markov chains segmentation for volume determination and quantitation in PET," *Phys. Med. Biol.*, vol. 52, pp. 3467–3491, 2007.

[12] S. Kamata, R. O. Eason, and Y. Bandou, "A new algorithm for N-dimensional Hilbert scanning," *IEEE Trans. Image Process.*, vol. 8, no. 7, pp. 964–973, Jul. 1999.

[13] Y. Delignon, A. Marzouki, and W. Pieczynski, "Estimation of generalized mixtures and its application in image segmentation," *IEEE Trans. Image Process.*, vol. 6, no. 10, pp. 1364–1375, Oct. 1997.

[14] D. Visvikis, A. Turzo, S. Gouret, P. Damine, F. Lamare, Y. Bizais, and C. Cheze Le Rest, "Characterisation of SUV accuracy in FDG PET using 3-D RAMLA and the Philips Allegro PET scanner," *J. Nucl. Med.*, vol. 45, no. 5, p. 103P, 2004.

[15] H. Caillol, W. Pieczynski, and A. Hillon, "Estimation of fuzzy Gaussian mixture and unsupervised statistical image segmentation," *IEEE Trans. Image Process.*, vol. 6, no. 3, pp. 425–440, Mar. 1997.

[16] F. Salzenstein and W. Pieczynski, "Parameter estimation in hidden fuzzy Markov random fields and image segmentation," *Graphical Models Image Process.*, vol. 59, no. 4, pp. 205–220, 1997.

[17] N. L. Johnson and S. Kotz*, Distributions in Statistics: Continuous Univariate Distributions*. New York: Wiley, 1970, vol. 1.

[18] G. Celeux and J. Diebolt, "L'algorithme SEM: Un algorithme d'apprentissage probabiliste pour la reconnaissance de mélanges de densités," *Revue de Statistique Appliquée*, vol. 34, no. 2, 1986.

[19] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Stat. Soc. B*, vol. 39, pp. 1–38, 1977.

[20] J. McQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symp. Math. Stat. Probabil.*, 1967, vol. 1, pp. 281–297.

[21] J. C. Dunn, "A fuzzy relative of the isodata process and its use in detecting compact well-separeted clusters," *J. Cybernet.*, vol. 31, pp. 32–57, 1974.

[22] K. Jordan, "IEC emission phantom appendix performance evaluation of positron emission tomographs," *Med. Public Health Res. Programme Eur. Commun.*, 1990.

[23] M. A. Lodge, V. Dilsizian, and B. R. Line, "Performance assessment of the Philips GEMINI PET/CT scanner," *J. Nucl. Med.*, vol. 45, p. 425, 2004.

[24] W. P. Segars, D. S. Lalush, and B. M. W. Tsui, "Modelling respiratory mechanics in the MCAT and spline-based MCAT phantoms," *IEEE Trans. Nucl. Sci.*, vol. 48, no. 1, pp. 89–97, Feb. 2001.

[25] C. D. Ramos, Y. E. Erdi, M. Gonen, E. Riedel, H. W. Yeung, H. A. Macapinlac, R. Chisin, and S. M. Larson, "FDG-PET standardized uptake values in normal anatomical structures using iterative reconstruction segmented attenuation correction and filtered back-projection," *Eur. J. Nucl. Med.*, vol. 28, pp. 155–164, 2001.

[26] F. Lamare, A. Turzo, Y. Bizais, C. Cheze-Le Rest, and D. Visvikis, "Validation of a Monte Carlo simulation of the Philips Allegro/Gemini PET systems using GATE," *Phys. Med. Biol.*, vol. 51, pp. 943–962, 2006.

[27] M. Hatt, P. Bailly, A. Turzo, C. Roux, and D. Visvikis, "PET functional volume segmentation: A robustness study," in *IEEE NSS-MIC Conf. Rec.*, 2008, pp. 4335–4339.

[28] N. Boussion, M. Hatt, and D. Visvikis, "Partial volume correction in PET based on functional volumes," *J. Nucl. Med.*, vol. 49, no. S1, p. 388, 2008.

Disponible en ligne sur ScienceDirect www.sciencedirect.com

Elsevier Masson France EM|consulte www.em-consulte.com

IRBM

IRBM 30 (2009) 144–149

Communication brève

# Une nouvelle méthode de détermination automatique des volumes fonctionnels pour les applications de l'imagerie d'émission en oncologie

## *A new automatic methodology for functional volumes determination in emission imaging for oncology applications*

M. Hatt [a,*], C. Cheze-Le Rest [a,c], A. Dekker [d], D. De Ruysscher [d], M. Oellers [d], P. Lambin [d], C. Roux [a,b], D. Visvikis [a]

[a] *U650 Inserm, laboratoire de traitement de l'information médicale (LaTIM), CHU Morvan, bâtiment 2Bis (I3S), 5, avenue Foch, 29609 Brest, France*
[b] *Telecom Bretagne, Institut Telecom, Brest, France*
[c] *Service de médecine nucléaire, CHU Morvan, Brest, France*
[d] *Radiation oncology department, MAASTRO clinic, 12, 6229 ET Maastricht, Pays-Bas*

Reçu le 15 avril 2009 ; accepté le 11 mai 2009
Disponible sur Internet le 1 juillet 2009

**Résumé**

La détermination des volumes fonctionnels est une étape cruciale pour les applications en oncologie comme le suivi thérapeutique ou la planification en radiothérapie guidée par l'image. Il n'existe pour l'instant pas de consensus dans la communauté sur la méthode appropriée pour définir automatiquement un volume tumoral sur l'image fonctionnelle d'émission (e.g. TEP au 18F-FDG), à cause de la grande variabilité des images obtenues dans ce contexte, en termes de bruit, de textures, de contrastes ou des formes et des fixations hétérogènes des tumeurs. Nous proposons une méthode automatique dont la robustesse et la précision ont été validées sur des acquisitions de fantôme, des tumeurs simulées et réelles, avec des performances très supérieures aux méthodes de référence par seuillage, constituant un outil prometteur pour les applications de la TEP en oncologie.


*Mots clés :* Oncologie ; Imagerie d'émission ; TEP ; Volumes fonctionnels ; Segmentation automatique

**Abstract**

Functional volumes determination is a crucial step for several applications in oncology like therapy assessment or image-guided radiotherapy treatment planning. There is currently no consensus about the appropriate method for an automatic definition of the tumoural volume on functional emission images (e.g. 18F-FDG PET), because they are characterized by a large variability of noise, textures and contrasts, as well as shapes and uptakes of tumours. We propose a novel automatic method that was validated for robustness and accuracy on phantom acquisitions, realistic simulations and clinical images of complex tumours. This method outperforms the reference thresholding methodologies and may have an impact in several PET applications in oncology.


*Keywords:* Oncology; Emission imaging; PET; Functional volumes; Automatic segmentation

## 1. Introduction

La détermination des volumes fonctionnels est une étape cruciale pour les applications en oncologie comme le suivi thérapeutique [1] ou la planification en radiothérapie guidée par l'image [2]. Cette tâche est généralement effectuée à la main par les utilisateurs, a été jugée complexe et est associée à une

* Auteur correspondant.
*Adresse e-mail :* hatt@univ-brest.fr (M. Hatt).

1959-0318/$ – see front matter © 2009 Elsevier Masson SAS. Tous droits réservés.
doi:10.1016/j.irbm.2009.05.004

très grande variabilité inter- et intra-utilisateurs [3]. Malgré le grand nombre de méthodes ayant été proposées récemment pour automatiser la définition des volumes tumoraux sur les images d'émission [4–19], il n'existe pour l'instant pas de consensus dans la communauté sur la méthode appropriée pour définir automatiquement un volume tumoral sur l'image fonctionnelle d'émission (e.g. TEP au 18F-FDG), du fait de leur qualité limitée et de la faiblesse de la plupart des méthodes proposées jusqu'à présent, inappropriées pour gérer la grande variabilité des paramètres en termes de bruit, de flou, de contraste ou des formes et des fixations hétérogènes des tumeurs. La plupart des méthodes proposées jusqu'à présent partagent également le désavantage de ne pas bénéficier d'une validation suffisante pour convaincre les utilisateurs finaux, par exemple en se contentant de résultats sur fantômes simplifiés ou sur des données cliniques sans vérité terrain.

Nous proposons une méthode automatique, robuste et précise, validée sur plusieurs ensembles de données de la simulation à l'image clinique, pouvant être utilisée sans optimisation préalable pour un scanner et un protocole spécifique et limitant l'intervention de l'utilisateur à la détection de la tumeur.

## 2. Matériels et méthodes

La méthode proposée est fondée sur l'utilisation du contexte méthodologique de la segmentation statistique d'images [20]. Plusieurs approches ont déjà tenté d'utiliser ce type de méthodologies dans le cadre de l'imagerie d'émission [14,15] mais ces dernières se limitaient à l'utilisation d'une mesure statistique dite « dure » où seule une somme de Dirac est considérée. Cette mesure permet de bien modéliser l'aspect bruité des images, mais n'est pas adaptée aux images d'émission qui sont de plus floues. En effet, l'hypothèse de la modélisation « dure » consiste à considérer qu'un voxel appartient à une classe et que son observation est bruitée, en conséquence de quoi elle ne permet pas de modéliser qu'un voxel puisse contenir un mélange de classes. La modélisation que nous utilisons est fondée sur l'utilisation d'une mesure statistique définie par un mélange de masses de Dirac pour les classes homogènes et de mesures continues de Lesbegue pour les transitions floues entre les régions [21]. Ceci permet de prendre en compte simultanément les deux principaux défauts des images d'émission : le bruit statistique et le flou induit par la résolution spatiale.

Cette mesure floue a été utilisée dans le cadre d'une modélisation par chaînes de Markov [19] puis d'une approche locale adaptative [22] offrant des performances supérieures. Cette dernière approche a été nommée FLAB pour *fuzzy locally adaptive Bayesian* et a fait l'objet d'un brevet[1]. De plus, en étendant la modélisation à trois classes homogènes et trois transitions floues différentes, la méthode est capable de prendre en compte l'hétérogénéité de la fixation au sein des tumeurs et permet de générer des volumes segmentés non binaires [23]. Cela est notamment intéressant pour les applications de « *dose painting* » en radiothérapie pour une optimisation de la dosimétrie [24] alors que la plupart des méthodes proposées jusqu'à présent ne considèrent que des segmentations binaires des volumes fonctionnels et ne peuvent donc pas être utilisées automatiquement dans cette optique. Dans la méthode FLAB, le contexte spatial des voxels est pris en compte par un cube glissant, au sein de l'estimation itérative des paramètres d'intérêt du modèle, à savoir les moyennes et variances de chaque classe ainsi que les probabilités a priori de chaque voxel d'appartenir à une classe donnée [22], ce qui est indispensable pour une segmentation précise. Cette estimation itérative est réalisée grâce à l'algorithme *stochastic expectation maximization* (SEM) [25] qui assure une vitesse de convergence supérieure et une relative indépendance aux conditions d'initialisation par rapport à l'algorithme *expectation maximization* (EM) classique.

Les résultats de FLAB ont été comparés avec ceux obtenus par des méthodes utilisant un seuillage fixe (ici 42 % du maximum comme proposé par Erdi et al. [4]) ou adaptatif prenant en compte le signal du fond environnant [9] dont les paramètres ont été optimisés pour les scanners considérés. Sa robustesse a été étudiée sur des acquisitions réelles de fantôme contenant des sphères homogènes et réalisées sur plusieurs scanners différents (Philips Gemini et Gemini TF, Siemens Biograph, GE Discovery LS) avec leurs algorithmes de reconstruction (RAMLA, TF ML-EM, OSEM) (Fig. 1). Sa précision a été validée sur 20 images de tumeurs réalistes tant en termes de formes que de fixations, basées sur des acquisitions réelles de patients et simulées à l'aide de *Geant4 application for tomography emission* (GATE) [26,27] (Fig. 2), ainsi que sur un ensemble de 18 images de tumeurs pulmonaires réelles (Fig. 3). Concernant ces dernières, tous les patients ont été opérés et le diamètre maximal des tumeurs a été mesuré lors de l'étude macroscopique des pièces opératoires [28]. Le diamètre maximal des tumeurs déterminé par le pathologiste a été comparé à celui mesuré sur les volumes segmentés par chaque méthode considérée.

Il est important de noter que la méthode proposée n'est pas conçue pour être appliquée à l'image corps entier du patient car l'objectif n'est pas de détecter la tumeur, mais de la segmenter avec la plus grande précision possible. Elle est appliquée à une sélection contenant toute la tumeur, détectée et sélectionnée par l'utilisateur. Pour l'instant, le choix d'utiliser la méthode binaire (deux classes dures et une transition floue) ou la méthode à trois classes (trois classes dures et trois transitions floues) repose sur l'utilisateur en fonction de son appréciation de l'hétérogénéité de la tumeur à segmenter, mais il est possible d'automatiser cette initialisation, par exemple avec un algorithme de K-moyennes flou avec sélection automatique du nombre de classes par minimisation de l'entropie de l'histogramme de l'image comme proposé par Provost dans sa thèse [29].

## 3. Résultats

Les performances de FLAB, tant en termes de robustesse que de précision, sont largement supérieures à celles des méthodes de référence utilisant des seuillages. L'évaluation de la robustesse [30] est importante, car elle permet de déterminer si une méthode donnée peut être utilisée sur des images obtenues avec n'importe quel scanner sans optimisation préalable, contrairement aux

[1] Brevet français référence FR08/56089.

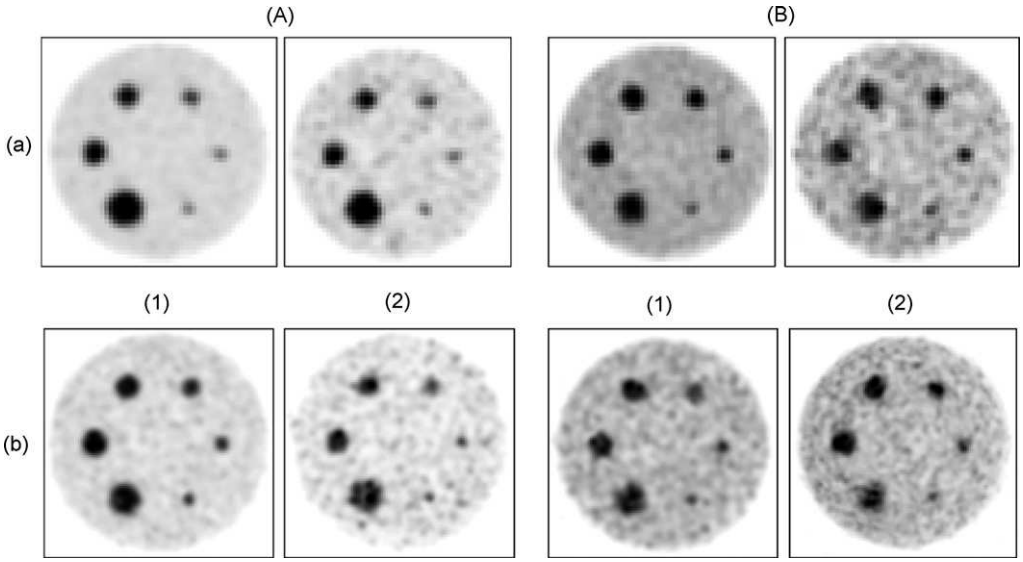



Fig. 1. Exemples d'acquisitions de fantôme sur un des scanners considérés (ici le Philips Gemini) avec différents paramètres. A. Contraste 8:1. B. Contraste 4:1 ; a : reconstruction 4 $mm^3 \times 4\,mm^3 \times 4\,mm^3$ ; b : reconstruction 2 $mm^3 \times 2\,mm^3 \times 2\,mm^3$ ; 1 : cinq minutes d'acquisition ; 2 : une minute d'acquisition.

méthodes utilisant des seuillages adaptatifs, dont les paramètres doivent être optimisés pour un scanner et une reconstruction donnés. Comme le montrent sur la Fig. 4, l'erreur moyenne et l'écart-type obtenus sur les sphères de 37 à 13 mm de diamètre, sur l'ensemble des acquisitions, FLAB permet d'obtenir moins de 10 % d'erreur sur les sphères, avec un écart-type de l'ordre de 5 à 10 %. Aucune des méthodes ne permet d'obtenir de bons résultats sur la sphère de 10 mm car on atteint ici les limites des scanners TEP dont la résolution spatiale est limitée à environ 5 mm de largeur à mi-hauteur, combinée à un échantillonnage spatial de voxels de 2 et 5 mm de côté. Les résultats sur fantômes sont satisfaisants pour démontrer la robustesse de la méthode et son universalité car elle donne de bons résultats sur différents types de textures, de bruit, d'échantillonnage spatial ou de contrastes. Bien que les erreurs soient de plus faibles concernant la détermination du volume (autour ou inférieures à 10 %), il ne s'agit que de cas idéaux de fixations sphériques uniformes sur un fond uniforme. Les tumeurs réelles présentent en règle générale des structures plus complexes, tant en termes de formes que d'hétérogénéité de fixations.

Ainsi, les résultats obtenus sur les tumeurs simulées (Fig. 5) permettent d'apprécier la précision de la segmentation en situa-

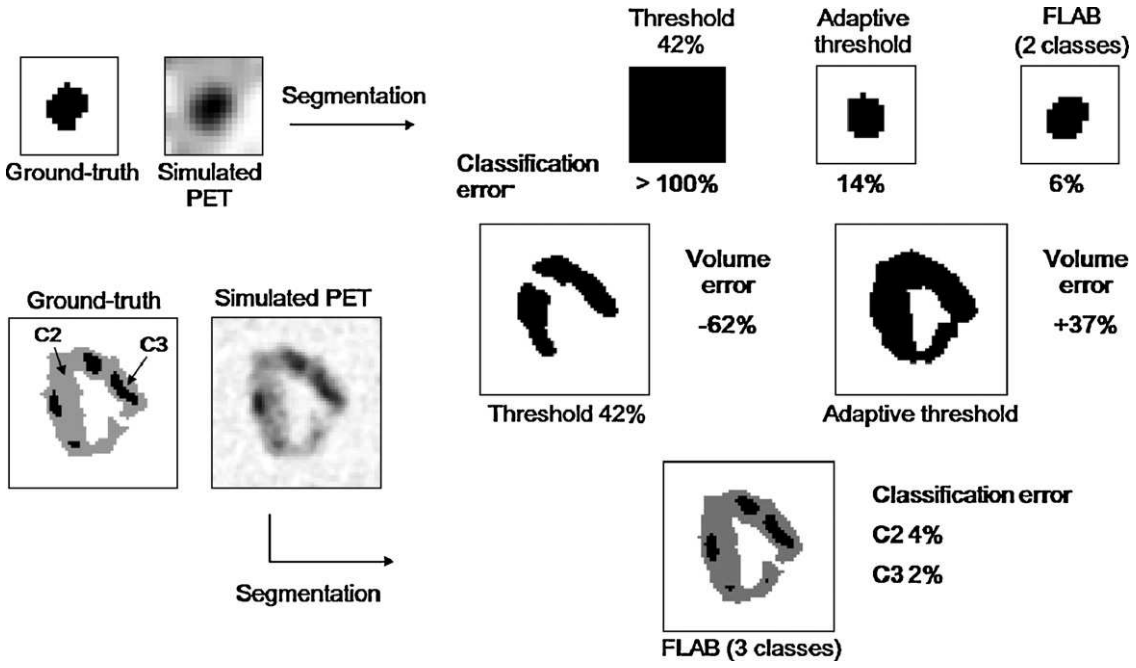



Fig. 2. Illustration de deux tumeurs simulées et des résultantes de segmentation obtenues par les différentes approches, par rapport à la vérité terrain simulée. L'échec complet du seuillage fixe à 42 % sur la plus petite tumeur s'explique par le fait que dans l'image, aucun voxel n'a de valeur inférieure à 42 % du maximum de la lésion (très faible contraste).

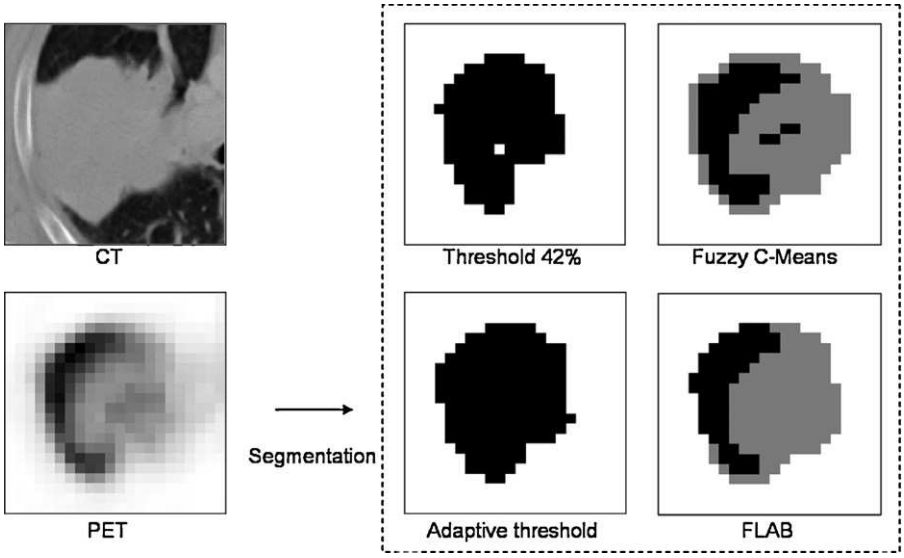



Fig. 3. Illustration d'une tumeur pulmonaire réelle (image anatomique et fonctionnelle) et les résultats de segmentation obtenus par différentes méthodes.

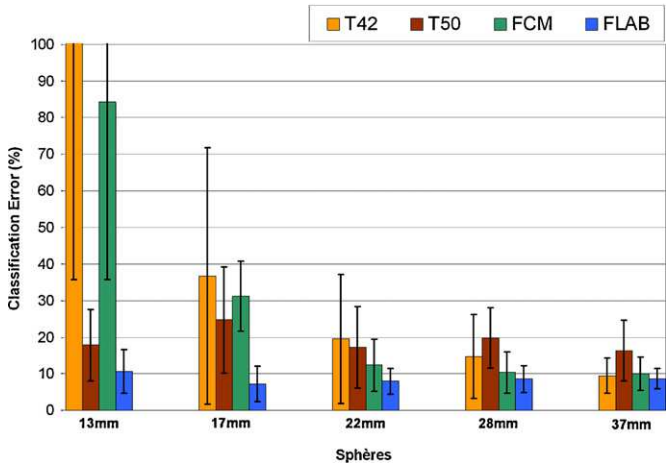



Fig. 4. Erreurs de classification voxel à voxel par rapport à la vérité terrain obtenues par différentes méthodes (seuillages à 42 et 50 % du maximum, T42 et T50, le *clustering* par Fuzzy C-Means FCM et FLAB) sur l'ensemble des acquisitions de fantôme (tous les scanners, tous les paramètres considérés).

tion plus réaliste. Les résultats sont en faveur de FLAB, avec une erreur moyenne (calculée sur un ensemble de 20 tumeurs) de classification voxel à voxel par rapport à la vérité terrain simulée inférieure à 9 % et un écart-type de 8 %. Le seuillage

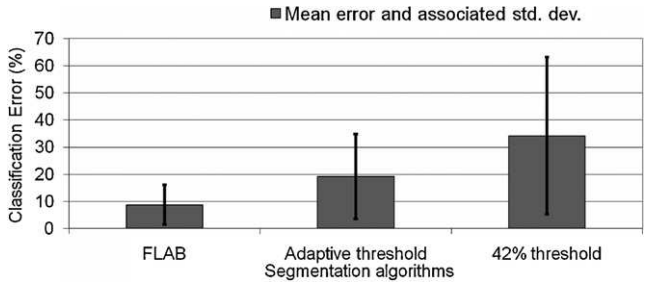



Fig. 5. Erreurs de classification moyenne et écart-type de chaque méthode sur 20 tumeurs simulées par rapport à la vérité terrain voxel à voxel.

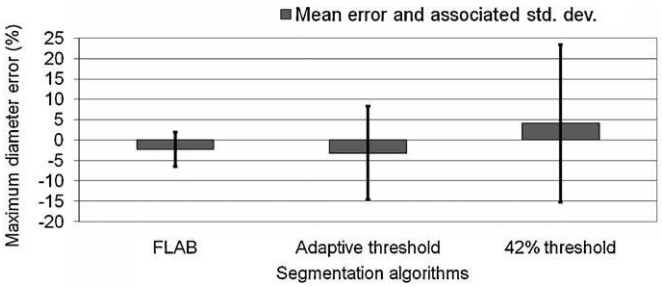



Fig. 6. Erreur moyenne et écart-type par rapport au diamètre de la tumeur mesuré en histologie, pour l'ensemble des 18 tumeurs et les différentes approches de segmentation des images.

adaptatif et le seuillage à 42 % donnent respectivement des erreurs moyennes et des écarts-types de $19 \pm 15$ % et $34 \pm 20$ %. Les résultats obtenus sur la mesure du diamètre maximal des tumeurs réelles mesurées en histologie (Fig. 6) sont également en faveur de FLAB. En effet, bien que toutes les méthodes obtiennent une erreur moyenne inférieure à 5 %, les écarts-types associés au seuillage adaptatif et au seuillage fixe sont respectivement de 10 et 20 % là où celui associé à FLAB est inférieur à 5 %. La faible erreur moyenne obtenue s'explique par le fait que dans les 18 tumeurs considérées, environ la moitié est sous-estimée et l'autre moitié surestimée. De plus, outre une précision accrue sur la définition des volumes tumoraux, comme l'illustrent les Fig. 2 et 3, FLAB est capable de générer des volumes segmentés non binaires, offrant une information supplémentaire très importante, notamment en radiothérapie, sur l'éventuelle hétérogénéité de la fixation au sein des tumeurs considérées.

## 4. Discussion et conclusion

La position de la TEP comme outil de référence pour le diagnostic en oncologie a été renforcée par l'arrivée des scanners multimodalités TEP/TDM depuis le début des années 2000. Plus

récemment, l'intérêt porté à l'imagerie fonctionnelle pour des applications comme le suivi thérapeutique et surtout la planification de traitement en radiothérapie par l'image a motivé le développement par de nombreux groupes de recherche, de méthodes permettant d'améliorer la détermination automatique des volumes fonctionnels. Il a déjà été montré que l'utilisation de l'imagerie fonctionnelle dans le cadre de la radiothérapie permet, d'une part, de réduire la variabilité inter- et intra-utilisateurs [28,31] et, d'autre part, d'inclure des volumes tumoraux qui sont ratés, par l'utilisation de l'imagerie anatomique seule, ou au contraire d'exclure des volumes non malins qui auraient été inclus à tort en se basant uniquement sur l'information de densité de tissus fournie par l'imagerie TDM [32].

La méthode que nous proposons a plusieurs avantages par rapport aux méthodes de référence utilisant des seuillages. Elle est d'abord plus robuste et peut être utilisée sur des images acquises sur différents scanners et reconstruites avec différents algorithmes, sans optimisation préalable de paramètres. La dépendance au scanner et aux caractéristiques de l'image est donc réduite par rapport aux seuillages adaptatifs. La précision de la méthode est supérieure, y compris et surtout sur des cas complexes de tumeurs hétérogènes sur lesquelles les méthodes binaires utilisant des seuillages sont inappropriées et échouent parfois totalement. La possibilité de générer directement des volumes segmentés à trois classes permet d'envisager l'implémentation automatique du principe de *dose painting* en radiothérapie, pour une dosimétrie optimisée, ou une analyse fine région par région de la tumeur dans le cadre du suivi thérapeutique. Enfin, elle est automatique et réduit l'intervention de l'utilisateur à la détection de la tumeur et son isolation dans une boîte de traitement. La méthode a été validée à la fois sur de multiples acquisitions de fantômes pour valider sa reproductibilité et sa robustesse et sur des images simulées et réelles de tumeurs complexes et hétérogènes pour valider sa précision. Les résultats encourageants obtenus par cette approche permettent de penser qu'il s'agit là d'une méthode pouvant avoir un impact important dans les diverses applications de la TEP : le diagnostic, le suivi thérapeutique et la radiothérapie, pour lesquelles une définition automatique et précise des volumes fonctionnels permet d'améliorer et d'accélérer l'analyse quantitative des images d'émission. Une étude est en cours dans le cadre d'un projet ANR (SIFR, 2009–2010) pour renforcer la validation de FLAB et estimer son impact dans le cadre de la radiothérapie guidée par l'image ainsi que pour le suivi thérapeutique.

## 5. Conflits d'intérêts

Aucun.

## Références


[1] Couturier O, Jerusalem G, N'Guyen JM, Hustinx R. Sequential positron emission tomography using (18F)fluorodeoxyglucose for monitoring response to chemotherapy in metastatic breast cancer. Clin Cancer Res 2006;12(21):6437–43.

[2] Jarritt H, Carson K, Hounsel AR, Visvikis D. The role of PET/CT scanning in radiotherapy planning. Br J Radiol 2006;79(S):27–35.

[3] NKrak NC, Boellaard R, et al. Effects of ROI definition and reconstruction method on quantitative outcome and applicability in a response monitoring trial. Eur J Nucl Med Mol Im 2005;32:294–301.

[4] Erdi YE, Mawlawi O, Larson SM, et al. Segmentation of lung lesion volume by adaptative positron emission tomography image thresholding. Cancer 1997;80(S12):2505–9.

[5] Greco C, Rosenzweig K, Cascini GL, et al. Current status of PET/CT for tumour volume definition in radiotherapy treatment planning for non-small cell lung cancer (NSCLC). Lung Cancer 2007;57(2):125–34.

[6] Nestle U, Kremp S, Schaefer-Schuler A, et al. Comparison of different methods for delineation of 18F-FDG PET-positive tissue for target volume definition in radiotherapy of patients with non-small cell lung cancer. J Nucl Med 2005;46(8):1342–8.

[7] Black QC, Grills IS, et al. Defining a radiotherapy target with positron emission tomography. Int J Rad Onco Bio Phys 2004;60:1272–82.

[8] Davis JB, Reiner B, et al. Assessment of 18(F) PET signals for automatic target volume definition in radiotherapy treatment planning. Rad Onco 2006;80:43–50.

[9] Daisne JF, Sibomana M, Bol A, et al. Tridimensional automatic segmentation of PET volumes based on measured source-to-background ratios: influence of reconstruction algorithms. Rad Onco 2003;69:247–50.

[10] Van Dalen J, Hoffman AL, et al. A novel iterative method for lesion delineation and volumetric quantification with FDG PET. Nucl Med Commun 2007;28:485–93.

[11] White MB. A semiautomatic approach to the delineation of tumour boundaries from PET data using Level Sets. Society of Nuclear Medicine annual meeting 2005; 314.

[12] Tylski P, Bonniaud G, Decenciere E, et al. 18F-FDG PET images segmentation using morphological watershed: a phantom study. IEEE NSS-MIC conference records 2006:2063–7.

[13] Zhu W, Jiang T. Automation segmentation of PET image for brain tumours. IEEE NSS-MIC conference records 2003;4:2627–9.

[14] Montgomery DWG, Amira A, Zaidi H. Fully automated segmentation of oncological PET volumes using a combined multiscale and statistical model. Medical Physics 2007;34(2):722–36.

[15] Demirkaya O. Lesion segmentation in whole body images of PET. IEEE NSS-MIC conference records 2003:2873–6.

[16] Geets X, Lee JA, Bol A, et al. A gradient-based method for segmenting FDG-PET images: methodology and validation. Eur J Nuc Med Mol Im 2007;34:1427–38.

[17] Li H, Thorstad WL, Biehl KJ, et al. A novel PET tumor delineation method based on adaptive region-growing and dual-front active contours. Medical Physics 2008;35(8):3711–21.

[18] Yu H, Caldwell C, Mah K, et al. Coregistered FDG PET/CT-based textural characterization of head and neck cancer for radiation treatment planning. IEEE Trans Med Imag 2008;28(3):374–83.

[19] Hatt M, Lamare F, Boussion N, et al. Fuzzy hidden Markov chains segmentation for volume determination and quantitation in PET. Phys Med Biol 2007;52:3467–91.

[20] Pieczynski W. Statistical image segmentation machine graphics and vision 1992; 1(1/2):261–8.

[21] Caillol H, Pieczynski W, Hillion A. Estimation of fuzzy Gaussian mixture and unsupervised statistical image segmentation. IEEE Trans Im Proc 1997;6(3):425–40.

[22] Hatt M, Turzo A, Roux C, Visvikis D. A fuzzy locally adaptive Bayesian segmentation approach for volume determination in PET. IEEE Trans Med Imag 2009;28(6):881–93.

[23] Hatt M, Dekker A, De Ruysscher D, Oellers M, Lambin P, Roux C, et al. Accurate functional volume definition in PET for radiotherapy treatment planning. IEEE NSS-MIC Conference records 2008:5567–71.

[24] Sovik A, Malinen E, Olsen DR. Strategies for biologic image-guided dose escalation: a review. Int J Rad Onco Biol Phys 2009;73(3):650–8.

[25] Celeux G, Diebolt J. L'algorithme SEM : un algorithme d'apprentissage probabiliste pour la reconnaissance de mélanges de densités. Rev Statist Appl 1986;34(2):35–52.

[26] Lamare F, Turzo A, Bizais Y, Cheze-Le Rest C, Visvikis D. Validation of a Monte Carlo simulation of the Philips Allegro/Gemini PET systems using GATE. Phys Med Biol 2006;51:943–62.

[27] Le Maitre A, Segars WP, Marache S, Reilhac A, Hatt M, Tomei S, et al. Incorporating patient specific variability in the simulation of realistic whole body 18F-FDG distributions for oncology applications. Proceedings of the IEEE (special issue on computational anthropomorphic anatomical models) 2009 [in press].

[28] Van Baardwijk A, Bosmans G, Boersma L, et al. PET-CT-based auto-contouring in non-small-cell lung cancer correlates with pathology and reduces interobserver variability in the delineation of the primary tumour and involved nodal volumes. Int J Rad Onco Biol Phys 2007;68(3):771–8.

[29] Provost JN. Classification bathymétrique en imagerie multispectrale SPOT. Brest: université de Bretagne occidentale; 2001 [thèse].

[30] Hatt M, Bailly P, Turzo A, Roux C, Visvikis D. PET functional volume segmentation: a robustness study. IEEE NSS-MIC conference records 2008:4335–9.

[31] Fox JL, Rengan R, O'meara W, et al. Does registration of PET and planning CT images decrease interobserver and intraobserver variation in delineating tumor volumes for non-small-cell lung cancer? Int J Rad Onco Biol Phys 2005;62(1):70–5.

[32] Ashamallaa H, Raa S, Parikh K, et al. The contribution of integrated PET/CT to the evolving definition of treatment volumes in radiation treatment planning in lung cancer. Int J Rad Onco Biol Phys 2005;63(4): 1016–23.

Int. J. Radiation Oncology Biol. Phys., Vol. 77, No. 1, pp. 301–308, 2010

0360-3016/10/$–see front matter

ELSEVIER

doi:10.1016/j.ijrobp.2009.08.018

**PHYSICS CONTRIBUTION**

# ACCURATE AUTOMATIC DELINEATION OF HETEROGENEOUS FUNCTIONAL VOLUMES IN POSITRON EMISSION TOMOGRAPHY FOR ONCOLOGY APPLICATIONS

Mathieu Hatt, Ph.D.,* Catherine Cheze le Rest, Ph.D., M.D.,*† Patrice Descourt, Ph.D.,* André Dekker, Ph.D.,‡ Dirk De Ruysscher, Ph.D., M.D.,‡ Michel Oellers, Ph.D.,‡ Philippe Lambin, Ph.D., M.D.,‡ Olivier Pradier, Ph.D., M.D.,*§ and Dimitris Visvikis, Ph.D.*

*Institut National de la Santé et de la Recherche Médicale U650 Brest, France; †Centre Hospitalier Universitaire, Morvan, Brest, France; ‡MAASTricht Radiation Oncology Clinic, Maastricht, The Netherlands; §Institute of Oncology, Centre Hospitalier Universitaire Morvan, Brest, France

**Purpose: Accurate contouring of positron emission tomography (PET) functional volumes is now considered crucial in image-guided radiotherapy and other oncology applications because the use of functional imaging allows for biological target definition. In addition, the definition of variable uptake regions within the tumor itself may facilitate dose painting for dosimetry optimization.**

**Methods and Materials: Current state-of-the-art algorithms for functional volume segmentation use adaptive thresholding. We developed an approach called fuzzy locally adaptive Bayesian (FLAB), validated on homogeneous objects, and then improved it by allowing the use of up to three tumor classes for the delineation of inhomogeneous tumors (3-FLAB). Simulated and real tumors with histology data containing homogeneous and heterogeneous activity distributions were used to assess the algorithm's accuracy.**

**Results: The new 3-FLAB algorithm is able to extract the overall tumor from the background tissues and delineate variable uptake regions within the tumors, with higher accuracy and robustness compared with adaptive threshold ($T_{bckg}$) and fuzzy C-means (FCM). 3-FLAB performed with a mean classification error of less than 9% ± 8% on the simulated tumors, whereas binary-only implementation led to errors of 15% ± 11%. $T_{bckg}$ and FCM led to mean errors of 20% ± 12% and 17% ± 14%, respectively. 3-FLAB also led to more robust estimation of the maximum diameters of tumors with histology measurements, with <6% standard deviation, whereas binary FLAB, $T_{bckg}$ and FCM lead to 10%, 12%, and 13%, respectively.**

**Conclusion: These encouraging results warrant further investigation in future studies that will investigate the impact of 3-FLAB in radiotherapy treatment planning, diagnosis, and therapy response evaluation.** © **2010 Elsevier Inc.**

**Heterogeneous functional volumes delineation, Automatic segmentation, Image-guided radiotherapy, Dose painting.**

## INTRODUCTION

Although most clinical applications of positron emission tomography (PET) rely on manual and visual analysis, accurate functional volume delineation in PET is crucial for numerous oncology applications. These include the use of tumor volume and associated determination of semiquantitative indices of activity concentration for diagnosis and therapy response evaluation (1) or the definition of target volumes in intensity-modulated radiation therapy (IMRT) (2). Subjective (1) and tedious manual delineation cannot perform accurate and reproducible segmentation, particularly when considering complex shapes and nonhomogeneous uptake. This results from the low quality of PET images due to statistical noise and partial volume effects (PVE) (3), arising from the scanner's limited spatial resolution.

Most of the previously proposed methods for PET volume definition are semiautomatic and threshold-based, using either fixed (30%–75% of the maximum activity) (2, 4, 5) or adaptive approaches incorporating the background activity (6–10). Unfortunately, these approaches often require additional *a priori* information and are user- and system-dependent. They require manual background regions of interest (ROIs), and their performance depends on parameters requiring optimization using phantom acquisitions for

Reprint requests to: Mathieu HATT, Ph.D., INSERM U650, LATIM, CHU MORVAN, 5 avenue Foch, 29609 Brest. Tel: (+33) 298018111; Fax: (+33) 298018124; E-mail: hatt@univ-brest.fr

Conflict of interest: none.

*Acknowledgments*—This work was supported by the Brittany Region grant program (Grant No. 1202-2004), the French National Research Agency (Grant Nos. ANR-06-CIS6-004-03 and ANR-08-ETEC-005-01), and Cancéropôle Grand Ouest (R05014NG).

Received Mar 31, 2009, and in revised form July 9, 2009. Accepted for publication Aug 13, 2009.

each scanner and reconstruction. Finally, all of these approaches are strictly binary and were not validated considering heterogeneous volumes.

Numerous works have addressed PET lesion segmentation using more advanced image segmentation methodologies (11–19). However, the majority of these approaches often depend on pre- or postprocessing steps such as deconvolution or denoising, are often binary only, and are validated on phantom acquisitions or clinical data without rigorous ground truth.

We previously developed an algorithm for PET volume definition by combining a fuzzy measure with a locally adaptive Bayesian-based classification (FLAB) that has been shown to perform better with respect to fixed thresholding, fuzzy C-means (FCM), or fuzzy hidden Markov chains (FHMC) for PET volume definition, as far as homogeneous spheres or slightly heterogeneous and nonspherical tumors are concerned (20). Preliminary results show that FLAB is also robust with respect to variability of the acquisition and reconstruction parameters (24).

Clinical tumors may be characterized by heterogeneous uptake, thus demanding a nonbinary approach for an accurate segmentation that may have a significant impact in defining biological target volumes for dose painting (21). The goals of this work were to (*1*) improve the FLAB model by incorporating the use of three hard classes and three fuzzy transitions and (*2*) evaluate its accuracy on real (with known diameter measured in histology) and simulated (with known ground truth) data sets containing inhomogeneous tumors.

## METHODS AND MATERIALS

### *Three-class fuzzy Bayesian segmentation (3-FLAB)*

The 3-FLAB algorithm is an extension of our previous work considering only a binary segmentation (20). FLAB automatically estimates parameters of interest from the image, maximizing the probability of each voxel to belong to one of the considered classes. This probability is estimated for each voxel as a function of its value and the values of its neighbors relative to the voxels' statistical distributions in the image, which corresponds to an estimation of the noise within each class. Hence, each voxel of the volume is considered a random variable within a Bayesian framework:

$$P(\mathrm{X}|\mathrm{Y}) = \frac{P(\mathrm{X},\mathrm{Y})}{P(\mathrm{Y})} = \frac{P(\mathrm{Y}|\mathrm{X})P(\mathrm{X})}{P(\mathrm{Y})}, \quad (1)$$

where $P(\mathrm{X}|\mathrm{Y})$is the probability of belonging to Class X knowing Observation Y. This probability is obtained by the product of $P(\mathrm{Y}|\mathrm{X})$ and $P(\mathrm{X})$, corresponding to the noise model and the spatial model, respectively. $P(\mathrm{Y}|\mathrm{X})$ is estimated considering the statistical distribution of the voxels within each class, whereas $P(\mathrm{X})$ is estimated using a sliding cube of $3 \times 3 \times 3$ voxels; hence, each voxel's classification is influenced by its neighbors. The parameters to estimate are the mean and variance of each class and the spatial probabilities of each voxel with respect to its neighbors. This is performed iteratively using a stochastic version (SEM) (25) of the Expectation Maximization (EM) (26) initialized with K-means (27) or fuzzy C-means (28). In addition, a fuzzy measure between the classes was added to account for the blur between regions, assuming each voxel may contain a mixture of classes (22, 23).

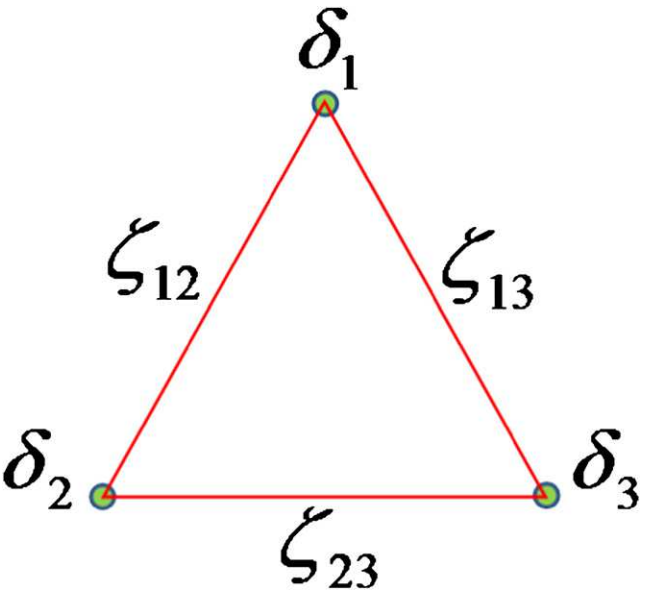



Fig. 1. The fuzzy scheme in the three-class fuzzy locally adaptive Bayesian (3-FLAB) implementation.

The difference between 3-FLAB and the previously developed binary-only FLAB (20) is the use of three classes and three fuzzy transitions within the model (see Fig. 1), to deal with both homogeneous and heterogeneous activity distributions. Figure 2 demonstrates the inability of FLAB to handle highly nonuniform activity distributions, where the lower uptake part of the lesion is erroneously considered as part of the background (see Fig. 2b), emphasizing the need to better model heterogeneous activity distributions. 3-FLAB should retain the accuracy and robustness of the original model, while also being able to handle the challenging heterogeneous activity distributions that are frequently seen in clinical lesions. The 3-FLAB segmentation workflow is summarized as follows, and the implementation and mathematical details can be found in the Appendix.

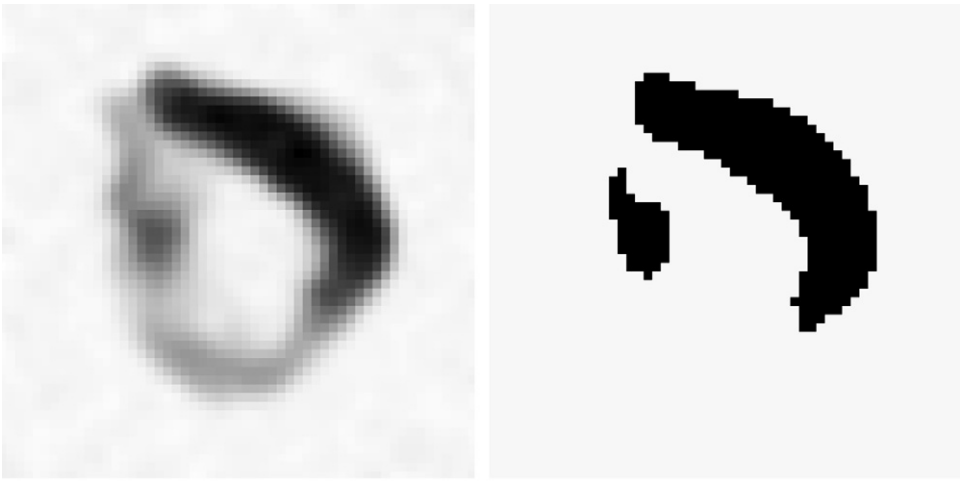
1. Initialization of both the spatial and noise models parameters: means and variances of each class are obtained using the K-means or fuzzy C-means. The prior probabilities are fixed at one third for each class.
2. Iterative estimation is performed using the SEM by stochastic sampling for each voxel according to its posterior probability.
3. Segmentation is done by selecting for each voxel the class or fuzzy level that maximizes its posterior probability and fusion of fuzzy levels with each hard class to generate a two- or three-class segmentation map.

### *Alternative segmentation methodologies used for comparison*

We compared the results of the 3-FLAB algorithm with the binary FLAB approach and the fuzzy C-means (with two or three clusters) clustering introduced by Dunn (28) and used to segment PET brain tumors in (13), as well as an adaptive thresholding (6) ($\mathrm{T_{bckg}}$):

$$I_{threshold} = \alpha \times I_{mean} + I_{background.} \quad (2)$$

$I_{mean}$ was obtained by computing the mean of all voxels contained inside an initial threshold at 70% of the maximum and $I_{background}$ by computing the mean of the voxels inside a ROI manually drawn on the background. $I_{mean}$ and $I_{background}$ were subsequently used to



(a) (b)

Fig. 2. Binary fuzzy locally adaptive Bayesian (FLAB) model applied to a heterogeneous simulated tumour (a). The segmentation result (b) clearly misses parts of the tumour.

derive a first approximation of the source-to-background contrast. The parameter $\alpha$ was optimized using phantom acquisitions on each scanner used to obtain the data. The adaptive thresholding algorithm was implemented using a region-growing approach with the maximum intensity voxel as a seed and iteratively adding three-dimensional (3D) neighboring voxels if their value was above the threshold calculated using Eq. 2.

### *Validation studies*

*Data sets:* Data Set 1 was used to evaluate the performance of the algorithm under realistic imaging conditions. It consists of 20 3D simulated tumors with variable levels of irregular shape and homogeneous or nonhomogeneous uptake distributions derived from tumors in patients undergoing 18F-fluorodeoxyglucose PET/CT investigations for radiotherapy treatment planning purposes. These images were acquired in 2D and 3D mode using the GE Discovery LS and Philips Gemini PET/CT scanners, respectively. Three of these tumors illustrating the range of sizes, shapes, and heterogeneities considered are shown in Fig. 4a–4c. The goal was to produce realistic images of PET tumors while retaining a voxel-based ground truth to compute accurate voxel-based classification errors. Half of the tumors were simulated considering a homogeneous uptake distribution, whereas the other half was simulated using significant heterogeneity within the tumor. The procedure followed to generate these images is illustrated in Fig. 3 and detailed in the following paragraphs.

Each clinical tumor is first manually delineated on the PET image by a nuclear medicine expert, thus creating a voxelized volume that represents the ground truth of the simulation. The activity levels attributed to each of the tumor parts were derived from the average activity measured in the same areas of the tumor in the corresponding patient images. This ground truth tumor structure is subsequently transformed into a nonuniform rational B-splines (NURBS) volume using Rhinoceros (CADLINK software, Morangis, France), for insertion into the NURBS-based CArdiac-Torso (NCAT) phantom (29) attenuation maps at the approximate position where it was located in the patient (30). No respiratory or cardiac motions were considered. Simulations using a model of the Philips PET/CT scanner previously validated with Geant4 Application for Tomography Emission (GATE) (31) were carried out. Forty-five million coincidences were simulated corresponding to the statistics of a clinical acquisition over a single-axial 18-cm field of view (31). Images were subsequently reconstructed using OPL-EM (seven iterations, one subset) (31) with two voxel sizes ($4 \times 4 \times 4$ for the Philips Gemini and $2 \times 2 \times 5$ mm$^3$ for the GE Discovery LS) to match those used in the corresponding clinical images.

Data Set 2 contains 18 images of lung tumors from patients with histologically proven non–small cell lung cancer (clinical Stage Ib–IIIb), acquired on the Siemens Biograph PET/CT scanner and reconstructed using OSEM (four iterations, eight subsets), with scatter and CT-based attenuation correction, and $5.31 \times 5.31 \times 3.38$ mm$^3$ voxels. These tumors were surgically extracted for a histology study in which their maximum diameter was measured by macroscopic examination (32). These diameters range from 15 to 90 mm ($44 \pm 21$). One of these tumors is shown in Fig. 4d.

*Analysis:* Because our goal is not the detection of a lesion in the whole image but the accurate estimation of its volume and shape, we assume it has been detected and isolated by the clinician within a 3D ''box'' encompassing the tumor.

Because a ground truth was available, classification errors (CE) were computed. In the case of a two-class ground truth, the CE is:

$$CE = \frac{card\{t|c_t \neq x_t\}}{card\{t|x_t = 1\}} \times 100, \quad (3)$$

where $c_t$ is the classification of voxel $t$, and $x_t$ is the true class. Card is the number of elements. This error measurement takes into consideration the spatial distribution of the tumor by considering both background voxels classified as object and object voxels classified as background. Consequently, this measure is more appropriate than simple volume estimation, which could lead to overall small volume errors associated with largely inaccurate segmentations. In addition, the errors are computed relatively to the size of the object, to avoid biases relative to the size of the processing box. In the case of a three-class ground truth, CE may be computed for each of the three classes using Eq. 4 or with respect to a binarized ground truth (second and third class merged) using Eq. 3.

$$CE_c = \frac{card\{t|x_t = c, c_t \neq c\} + card\{t|x_t \neq c, c_t = c\}}{card\{t|x_t = c\}} \times 100, \quad (4)$$

where $CE_c$ stands for the classification error associated with a given class c.

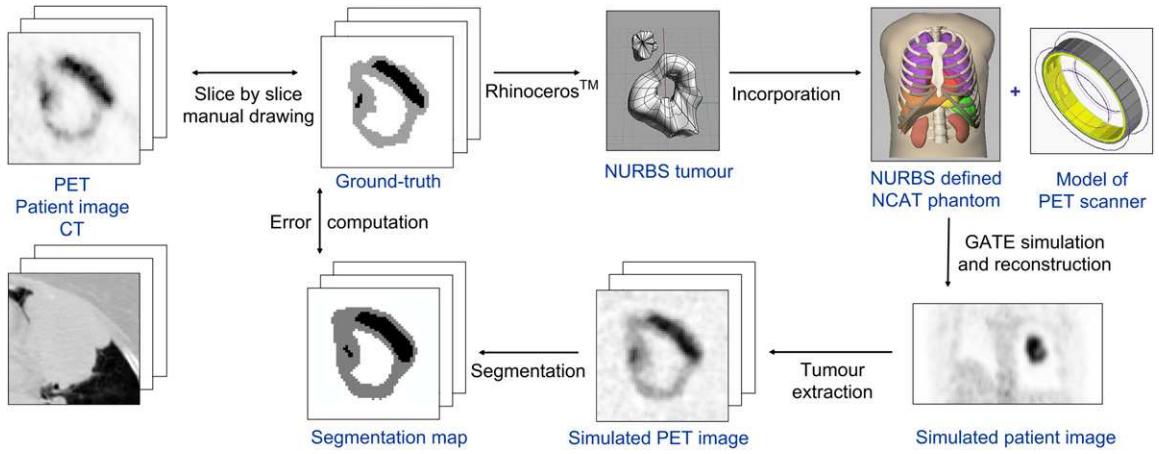



Fig. 3. The simulation of realistic positron emission tomography images.

Two analyses were conducted using Data Set 1. The first considered the entire data set (both homogeneous and heterogeneous tumors) and CE computed using Eq. 3 to compare overall performances of FLAB (binary only), 3-FLAB, FCM, and $T_{bckg}$. The second considered only the 10 heterogeneous tumors to compute $CE_2$ and $CE_3$ using Eq. 4 for 3-FLAB and FCM with three clusters.

The segmentation accuracy on the tumors with histology (Data Set 2) was assessed by segmenting the clinical image and subsequently measuring the maximum diameter on the segmented volumes to compare it with the histology measurement.

## RESULTS

Figure 5 contains one axial slice of the segmentations obtained on three simulated tumors of Data Set 1 and one tumor of Data Set 2. Figure 6a contains the mean classification errors and standard deviation obtained by all the methods on the 20 tumors of Data Set 1. FLAB (binary only) performed well on homogeneous tumors but failed as expected on strongly heterogeneous lesions, leading to overall errors of 15% ± 11%. 3-FLAB, in contrast, produced segmentation maps closer to the ground truth, both visually and quantitatively, with errors between 5% and 15% (9% ± 8%). FCM (with two or three clusters) was competitive with respect to 3-FLAB for some tumors but showed a higher variability (10%–40%) and mean error (20% ± 12%). This translated qualitatively in FCM being unable to differentiate two regions within the tumor as well as being unable to detect discontinuities in the contours (e.g., Fig. 5d, first row). In addition, for the regions where a transition was present between the high uptake region and the background (e.g., Fig. 4d), the 3-FLAB approach was the only one giving accurate representation of this transition (Fig. 5c vs. Fig. 5d, last row). $T_{bckg}$ was not able to produce satisfactory segmentation in several cases. Tumors with high overall contrast were approximately extracted from the background (e.g., Fig. 5e, rows 2–4). However, as a binary method, it is unable to delineate uptake distributions within the tumor. In several cases, the heterogeneity was significant, and $T_{bckg}$ lead to significant underevaluation of the tumor volume (CE up to 60% with a mean of 17% ± 14%) because it tends to extract the high-activity region or parts of the reduced uptake region only (e.g., Fig. 5e, first row).

Figure 6b compares 3-FCM (using three clusters) and 3-FLAB concerning the three-class segmentation of the 10 heterogeneous simulated tumors of Data Set 1. 3-FCM is less accurate and robust compared with 3-FLAB, especially in the delineation of higher activity regions (third class), with about twice the mean error and standard deviation (24% ± 20%) of 3-FLAB (11% ± 8%).

Figure 7 contains the mean error and standard deviation with respect to the maximum diameter, computed on the tumor histology database (Data Set 2). Whereas all methods gave relatively low mean errors (≤3%), the standard deviation associated with FCM and $T_{bckg}$ (13% and 12%, respectively) is about twice that of 3-FLAB (<6%), and binary FLAB showed a standard deviation of almost 10%. The low mean error for all these algorithms is explained by the fact that there were about the same amount of under- and overestimation of the diameters in this data set, resulting in an overall low mean error. Here the standard deviation is a better indicator of the accuracy obtained on the data set and demonstrates higher accuracy and robustness for 3-FLAB.

## DISCUSSION

Functional volume delineation represents an area of interest for multiple clinical applications (routine and research) of PET. Such areas include response to therapy studies and the use of biological tumor volumes in radiotherapy treatment planning. Although several fully automatic algorithms have recently been proposed (11–20), segmentation methodologies currently used in clinical practice are based on the use of fixed and adaptive thresholding (4–10). These algorithms have been shown to determine functional volumes accurately under specific imaging conditions of spherical and homogeneous activity distribution object in phantom studies and have been evaluated on clinical images for which the ground truth is unknown. In clinical practice, lesions are often heterogeneous in shape and uptake. To address these issues, we have extended a previously developed algorithm to evaluate

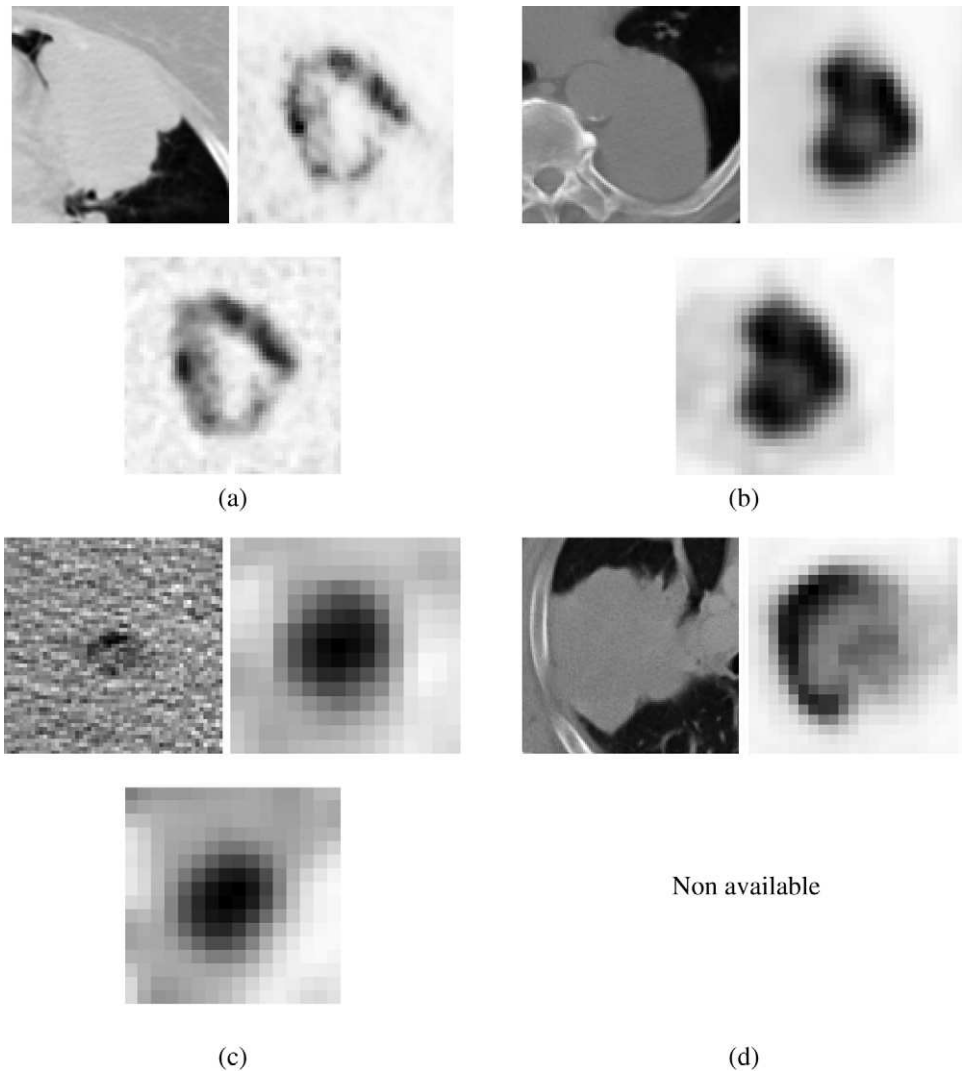



Fig. 4. Data sets illustration. (a–d) Examples of clinical tumors (up) with CT (left) and PET (right), and the corresponding simulated PET (down): (a–c) Data Set 1; (d) Data Set 2.

lesions with nonuniform uptake and nonspherical forms. In addition, we have proposed an evaluation framework including both realistic simulated patient lesions and histological assessment of tumor diameters, allowing for evaluation of segmentation algorithms under standard imaging conditions and the added advantage of knowing the ground truth.

The inability of the adaptive thresholding considered in this study to segment complex tumors accurately is demonstrated by its poor performance. This is explained by the fact that in cases of heterogeneous uptake, the 70% threshold used for the initial estimation of the tumor-to-background contrast may retain only the high uptake region, thus leading to incorrect contrast estimation. However, if the lesion is small or has a small contrast, the 70% threshold may lead to an initial overestimation of the volume of the tumor, and hence an underestimation of its uptake and an incorrect estimation of the contrast, for which the subsequent adaptive thresholding may not be able to compensate. In addition, the background ROI is user-dependent with a potentially high impact on the result, especially with heterogeneous background. In such cases, we systematically selected the ROI that resulted in the lowest error. Finally, the region growing implementation avoids incorporating false positives of the background if they are not connected to the main tumor, especially when the contrast is low or the background is noisy and heterogeneous. However, it also makes the algorithm dependent on the seed location and can lead to missing parts of the tumor when several high-uptake regions are connected by low-uptake regions. FCM can produce binary or three-class segmentations, but its robustness and accuracy are much lower compared with FLAB because it incorporates neither spatial correlation nor noise modeling. One advantage of the $T_{bckg}$ over FCM is its region growing implementation that makes it less susceptible than FCM to the inclusion of high-intensity voxels of the background. Therefore, FCM usually performs poorer than $T_{bckg}$ for low-contrast lesions and noisy images but better for heterogeneous activity distributions within the tumor. In contrast, 3-FLAB performed accurately even under challenging contrast, noise, and heterogeneity conditions, with

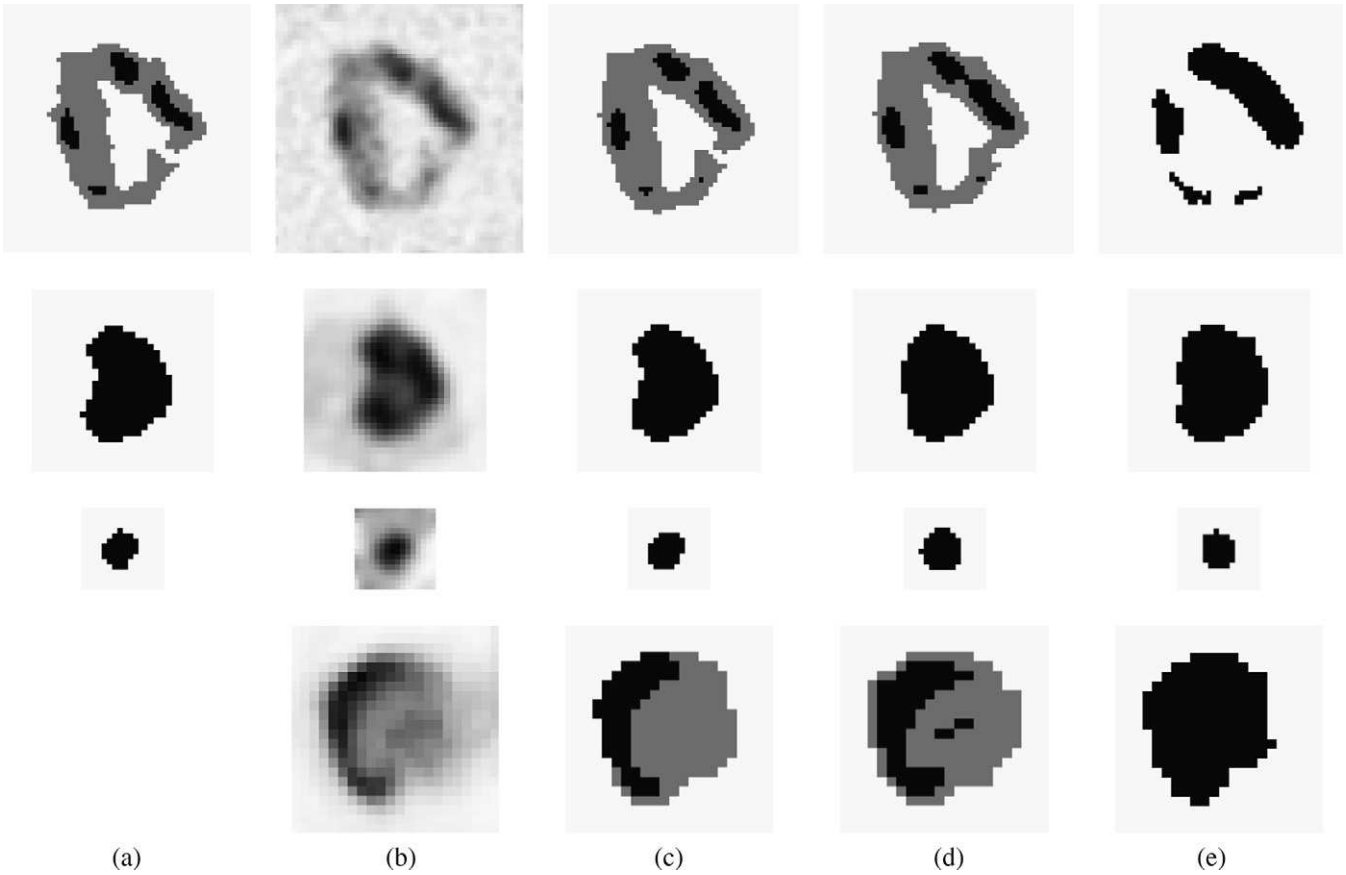


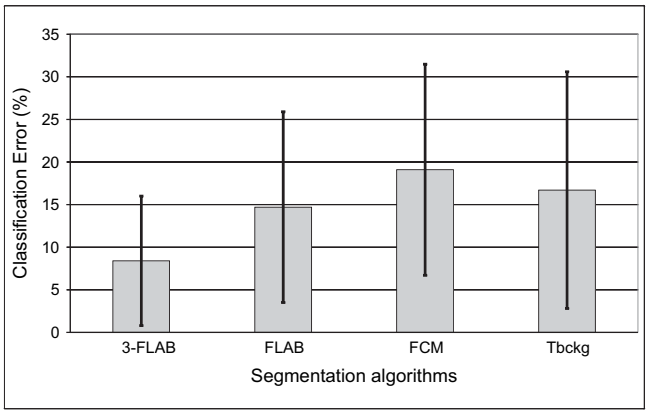

Fig. 5. Segmentations of the tumors in Fig. 4a–4d: (a) ground truth; (b) positron emission tomographic image; segmentations for (c) three-class fuzzy locally adaptive Bayesian, (d) fuzzy C-means, and (e) adaptive threshold models.

overall superior performance compared with the other algorithms considered here.

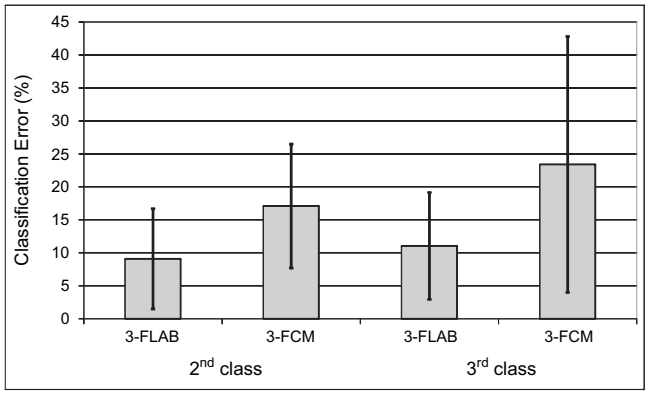
The need for more than three classes may arise for heterogeneous tumors on a heterogeneous background. However, all the clinical tumors considered in this study were correctly delineated using two or three classes because the contrasts between the heterogeneities within the tumor are usually much higher than those occurring in the background. Hence, only one hard class may be sufficient to deal with the background, whereas two are required to correctly handle the significantly different uptakes occurring inside the tumor. Eventually the 3-FLAB algorithm could be extended to more than three classes assuming that only pairs of hard classes generate fuzzy transitions. One also has to keep in mind that using more classes will lead to smaller regions, but those regions within the tumor will subsequently be used for quantification or radiotherapy dose boosting and/or painting and should therefore be kept reasonably large. The potential impact of using three classes proposed by 3-FLAB should therefore be investigated before more complex segmentations using additional classes can be considered.

We have already demonstrated that FLAB performs well for small lesions down to 13 mm in diameter (20), and this study was not designed to investigate specifically the ability of 3-FLAB to deal with small tumors because these rarely exhibit heterogeneous uptake that can be detected on the PET image considering the existing resolution limits. 3-FLAB retains all the characteristics of FLAB but also has the ability to consider a third class and therefore handle non-uniform lesion activity distributions. Thus, 3-FLAB does not as such improve the delineation of small (<2 cm) lesions. However, the higher/lower uptake regions within the larger tumors are often of small size, comparable to that of small lesions, with PVE affecting them with respect to their ''background,'' which is, in fact, the other part of the tumor with a different uptake. As Fig. 6b demonstrates, 3-FLAB is capable of accurately segmenting these regions.

An application that could greatly benefit from the use of FLAB is radiotherapy treatment planning (33). It is now acknowledged that planning based on PET/CT volumes improves tumor delineation by reducing inter- and intraobserver variability (32, 34). It can also lead to the inclusion of regions not visible on CT or the exclusion of regions without significant uptake (35). Using the 3-FLAB algorithm could help lower inter- and intraobserver variability, as well as shorten the time-consuming delineation process associated with currently implemented algorithms given the need for multiple phantom studies in the use of adaptive thresholding. 3-FLAB takes a few seconds per iteration even for the largest tumors considered in this study (on a single 2-Ghz core processor in C++ implementation). Further,




(a)




(b)

Fig. 6. Mean classification errors and standard deviation for (a) all methodologies considering all 20 tumors of Data Set 1, (b) three-class fuzzy locally adaptive Bayesian, and 3-class fuzzy C-means considering the second and third classes of the 10 heterogeneous tumors of Data Set 1.

''dose painting'' can be facilitated by the nonbinary nature of the proposed segmentation, allowing for automatic definition of ROIs inside the tumor—for example, in dose-escalation studies (36)—in addition to the external contour information for optimized dosimetry, potentially reducing the dose delivered to healthy surrounding tissues and organs. The impact of such improved accuracy on overall patient outcome remains to be demonstrated in clinical studies, which are planned for the future. Finally, FLAB robustness with respect to the noise characteristics associated with the use of different scanners, acquisition protocols, and reconstruction algorithms has been demonstrated in a preliminary study (24) and should allow its use with any type of PET images without the need for time-consuming preprocessing optimization.

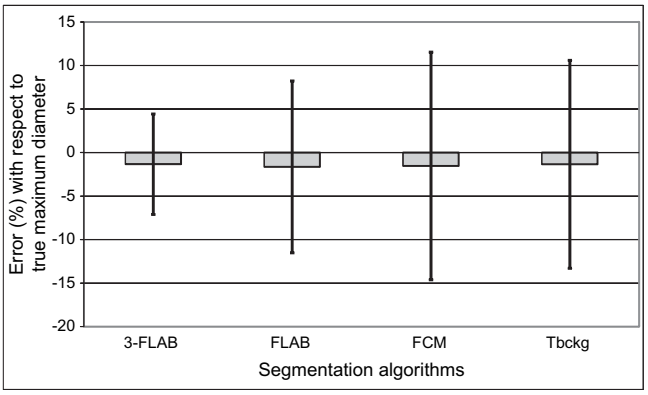



Fig. 7. Mean errors and standard deviation for each methodology, with respect to known maximum diameter of Data Set 2 tumors.

The proposed algorithm may also have an impact on diagnosis and therapy response assessment when combined with PVE correction (PVC) for accurate quantification. With various PVC approaches, anatomic information from MRI or CT is used to improve the quantitative and qualitative accuracy of functional images (37, 38). Unfortunately, when no anatomic image is available or no correlation exists between the anatomic and functional structures, such approaches are not easy to use (3). This is especially true in cases of large heterogeneous tumors for which there is little to no correlation between the anatomic and functional information. A potential solution will be the use of the FLAB result instead of the anatomic image in combination with one of the previously proposed PVC algorithms. This should lead to improved contrast at the object's borders as well as improved quantification in the regions within the tumor. Such combination recently demonstrated encouraging results (39) and warrants further investigation regarding the potential impact in clinical therapy response studies.

## CONCLUSION

A modified version of the FLAB algorithm has been developed to include the estimation of three hard classes and three fuzzy transitions. This automatic approach combines statistical and fuzzy modeling to address specific issues associated with 3D-PET images, such as noise and PVE. Its accuracy has been assessed on both simulated and clinical images of complex shapes containing inhomogeneous activities and small regions. The results demonstrate the ability of 3-FLAB to delineate such lesions, for which the threshold-based methodologies suggested until now have failed.

## REFERENCES

1. Krak NC, Boellaard R, Hoekstra OS, *et al*. Effects of ROI definition and reconstruction method on quantitative outcome and applicability in a response monitoring trial. *Eur J Nucl Med Mol Imaging* 2005;32:294–301.
2. Jarritt PH, Carson KJ, Hounsel AR, Visvikis D. The role of PET/CT scanning in radiotherapy planning. *Brit J Rad* 2006; 79(Suppl):S27–35.
3. Soret M, Bacharach SL, Buvat I. Partial-volume effect in PET tumour imaging. *J Nucl Med* 2007;48:932–945.
4. Erdi YE, Mawlawi O, Larson SW, *et al*. Segmentation of lung lesion volume by adaptive positron emission tomography image thresholding. *Cancer* 1997;80(Suppl 12):2505–2509.
5. Greco C, Rosenzweig K, Cascini GL, *et al*. Current status of PET/CT for tumour volume definition in radiotherapy treatment

planning for non-small cell lung cancer (NSCLC). *Lung Cancer* 2007;57:125–134.
6. Nestle U, Kremp S, Schaefer-Schuler A, *et al*. Comparison of different methods for delineation of 18F-FDG PET-positive tissue for target volume definition in radiotherapy of patients with non–small cell lung cancer. *Jour Nucl Med* 2005;46: 1342–1348.
7. Black QC, Grills IS, Kestin LL, *et al*. Defining a radiotherapy target with positron emission tomography. *Int J Radiat Oncol Biol Phys* 2004;60:1272–1282.
8. Davis JB, Reiner B, Huser M, *et al*. Assessment of 18(F) PET signals for automatic target volume definition in radiotherapy treatment planning. *Radiother Oncol* 2006;80: 43–50.
9. Daisne J-F, Sibomana M, Bol A, *et al*. Tri-dimensional automatic segmentation of PET volumes based on measured source-to-background ratios: influence of reconstruction algorithms. *Radiother Oncol* 2003;69:247–250.
10. Van Dalen JA, Hoffman AL, Dicken V, *et al*. A novel iterative method for lesion delineation and volumetric quantification with FDG PET. *Nucl Med Commun* 2007;28:485–493.
11. White CJ, Brady JM. A semi-automatic approach to the delineation of tumour boundaries from PET data using level sets. Society of Nuclear Medicine 52nd Annual Meeting. Toronto: Canada; June 2005. abstract 314.
12. Tylski P, Bonniaud G, Decenciere E, *et al*. 18F-FDG PET images segmentation using morphological watershed: A phantom study. *IEEE Neurosci Symp Conference Record* 2006;4: 2063–2067.
13. Zhu W, Jiang T. Automation segmentation of PET image for brain tumours. *IEEE Neurosci Symp Conference Record* 2003;4:2627–2629.
14. Montgomery DWG, Amira A, Zaidi H. Fully automated segmentation of oncological PET volumes using a combined multiscale and statistical model. *Med Phys* 2007;34: 722–736.
15. Demirkaya O. Lesion segmentation in wholebody images of PET. *IEEE Neurosci Symp Conference Record* 2003;4: 2873–2876.
16. Geets X, Lee JA, Bol A, *et al*. A gradient-based method for segmenting FDG-PET images: Methodology and validation. *Eur J Nucl Med Mol Imaging* 2007;34:1427–1438.
17. Li H, Thorstad WL, Biehl KJ, *et al*. A novel PET tumor delineation method based on adaptive region-growing and dual-front active contours. *Med Phys* 2008;35:3711–3721.
18. Yu H, Caldwell C, Mah K, *et al*. Co-registered FDG PET/CT-based textural characterization of head and neck cancer for radiation treatment planning. *IEEE Trans Med Imaging* 2009;28: 374–383.
19. Hatt M, Lamare F, Boussion N, *et al*. Fuzzy hidden Markov chains segmentation for volume determination and quantitation in PET. *Phys Med Biol* 2007;52:3467–3491.
20. Hatt M, Turzo A, Roux C, *et al*. A fuzzy Bayesian locally adaptive segmentation approach for volume determination in PET. *IEEE Trans Med Imaging* 2009;28:881–893.
21. Ling CC, Humm J, Larson S, *et al*. Towards multidimensional radiotherapy (MD-CRT): Biological imaging and biological conformality. *Int J Radiat Oncol Biol Phys* 2000;47:551–560.
22. Caillol H, Pieczynski W, Hillon A. Estimation of fuzzy Gaussian mixture and unsupervised statistical image segmentation. *IEEE Trans Image Processing* 1997;6:425–440.
23. Salzenstein F, Pieczynski W. Parameter estimation in hidden fuzzy Markov random fields and image segmentation. *Graphic Models Image Processing* 1997;59:205–220.
24. Hatt M, Turzo A, Bailly P, *et al*. Automatic delineation of functional volumes in PET: A robustness study. Presented at the Society of Nuclear Medicine 2009. Annual Meeting Toronto: Canada;June 1317.
25. Celeux G, Diebolt J. L'algorithme SEM: un algorithme d'apprentissage probabiliste pour la reconnaissance de mélanges de densités. *Revue Statistique Appliquée* 1986;34:35–52.
26. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *J R Stat Soc B* 1977;39: 1–38.
27. McQueen J. Some methods for classification and analysis of multivariate observations. *Proc 5th Berkeley Symp Math Stat Prob* 1967;1:281–297.
28. Dunn JC. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *J Cybernet* 1974; 31:32–57.
29. Segars WP. Development and application of the new dynamic NURBS-based cardiac-torso (NCAT) phantom [Ph.D. thesis]. Chapel Hill, NC: University of North Carolina; 2001.
30. Le Maitre A, Segars WP, Marache S, *et al*. Incorporating patient specific variability in the simulation of realistic whole body 18F-FDG distributions for oncology applications. *Proc IEEE*, in press.
31. Lamare F, Turzo A, Bizais Y, *et al*. Validation of a Monte Carlo simulation of the Philips Allegro/Gemini PET systems using GATE. *Phys Med Biol*, 2006;51:943–962.
32. Van Baardwijk A, Bosmans G, Boersma L, *et al*. PET-CT-based auto-contouring in non-small-cell lung cancer correlates with pathology and reduces interobserver variability in the delineation of the primary tumour and involved nodal volumes. *Int J Radiat Oncol Biol Phys* 2007;68:771–778.
33. Pan T, Mawlawi O. PET/CT in radiation oncology. *Med Phys* 2008;35:4955–4966.
34. Fox JL, Rengan R, O'Meara E, *et al*. Does registration of PET and planning CT images decrease interobserver and intraobserver variation in delineating tumor volumes for non-small-cell lung cancer? *Int J Radiat Oncol Biol Phys* 2005;62:70–75.
35. Ashamallaa H, Raa S, Parikh K, *et al*. The contribution of integrated PET/CT to the evolving definition of treatment volumes in radiation treatment planning in lung cancer. *Int J Radiat Oncol Biol Phys* 2005;63:1016–1023.
36. Sovik A, Malinen E, Olsen DR. Strategies for biologic image-guided dose escalation: A review. *Int J Radiat Oncol Biol Phys* 2009;73:650–658.
37. Rousset OG, Ma Y, Evans AC. Correction for partial volume effects in PET: Principle and validation. *J Nucl Med* 1998;39: 904–911.
38. Boussion N, Hatt M, Lamare F, *et al*. A multiresolution image based approach for correction of partial volume effects in emission tomography. *Phys Med Biol* 2006;51:1857–1876.
39. Boussion N, Hatt M, Visvikis D. Partial volume correction in PET based on functional volumes. *J Nucl Med* 2008;49(Suppl 1):388.

# Reproducibility of $^{18}$F-FDG and 3′-Deoxy-3′-$^{18}$F-Fluorothymidine PET Tumor Volume Measurements

Mathieu Hatt[1], Catherine Cheze-Le Rest[1,2], Eric O. Aboagye[3], Laura M. Kenny[3], Lula Rosso[3], Federico E. Turkheimer[3], Nidal M. Albarghach[1,4], Jean-Philippe Metges[4], Olivier Pradier[1,4], and Dimitris Visvikis[1]

*[1]INSERM, U650, LaTIM, CHU Morvan, Brest, France; [2]Academic Department of Nuclear Medicine, CHU Morvan, Brest, France; [3]MRC Clinical Sciences Centre, Imperial College London, Hammersmith Hospital, London, United Kingdom; and [4]Institute of Oncology, CHU Morvan, Brest, France*

The objective of this study was to establish the repeatability and reproducibility limits of several volume-related PET image–derived indices—namely tumor volume (TV), mean standardized uptake value, total glycolytic volume (TGV), and total proliferative volume (TPV)—relative to those of maximum standardized uptake value ($SUV_{max}$), commonly used in clinical practice. **Methods:** Fixed and adaptive thresholding, fuzzy C-means, and fuzzy locally adaptive Bayesian methodology were considered for TV delineation. Double-baseline $^{18}$F-FDG (17 lesions, 14 esophageal cancer patients) and 3′-deoxy-3′-$^{18}$F-fluorothymidine ($^{18}$F-FLT) (12 lesions, 9 breast cancer patients) PET scans, acquired at a mean interval of 4 d and before any treatment, were used for reproducibility evaluation. The repeatability of each method was evaluated for the same datasets and compared with manual delineation. **Results:** A negligible variability of less than 5% was measured for all segmentation approaches in comparison to manual delineation (5%–35%). $SUV_{max}$ reproducibility levels were similar to others previously reported, with a mean percentage difference of 1.8% ± 16.7% and −0.9% ± 14.9% for the $^{18}$F-FDG and $^{18}$F-FLT lesions, respectively. The best TV, TGV, and TPV reproducibility limits ranged from −21% to 31% and −30% to 37% for $^{18}$F-FDG and $^{18}$F-FLT images, respectively, whereas the worst reproducibility limits ranged from −90% to 73% and −68% to 52%, respectively. **Conclusion:** The reproducibility of estimating TV, mean standardized uptake value, and derived TGV and TPV was found to vary among segmentation algorithms. Some differences between $^{18}$F-FDG and $^{18}$F-FLT scans were observed, mainly because of differences in overall image quality. The smaller reproducibility limits for volume-derived image indices were similar to those for $SUV_{max}$, suggesting that the use of appropriate delineation tools should allow the determination of tumor functional volumes in PET images in a repeatable and reproducible fashion.

**Key Words:** oncology; PET; other; delineation; $^{18}$F-FDG; $^{18}$F-FLT; reproducibility; tumor volume

**J Nucl Med 2010; 51:1368–1376**
DOI: 10.2967/jnumed.110.078501

Received Apr. 28, 2010; revision accepted Jun. 10, 2010.
For correspondence or reprints contact: Mathieu Hatt, LaTIM, Inserm U650, CHU Morvan, 5 Ave. Foch, 29609 Brest, France.
E-mail: hatt@univ-brest.fr


Most current PET clinical practices for diagnosis, staging, prognosis, therapy-response assessment, and patient follow-up rely on manual and visual analysis (*1*). The index most commonly used in PET clinical studies is the standardized uptake value (SUV). To obtain this index of activity accumulation, a region of interest (ROI) should be determined, usually drawn manually or using some fixed threshold. Although an ROI is not the only factor that can affect the accuracy of SUVs, the type and size of an ROI are large contributors to the variability of such measurements, as has been previously demonstrated (*2,3*). A popular alternative is the use of the pixel with the maximum activity value, usually referred to as the maximum SUV ($SUV_{max}$). Many studies have demonstrated the prognostic and predictive value of $SUV_{max}$, despite the fact that it is sensitive to image noise (*4,5*). On the other hand, a few, mostly recent, studies have explored the use of overall tumor volume (TV) as an index for prognosis and response assessment (*6–8*). These studies considered the TV either alone or in combination with the mean SUV ($SUV_{mean}$), to form the total glycolytic volume (TGV) and total proliferative volume (TPV) (for $^{18}$F-FDG and 3′-deoxy-3′-$^{18}$F-fluorothymidine [$^{18}$F-FLT], respectively), defined as the product of TV × $SUV_{mean}$ (*9–11*).

The accuracy, robustness, repeatability, and reproducibility of image delineation are directly responsible for the reduced use of functional volumes derived from PET images. On the one hand, manual delineation of functional volumes using PET images leads to high inter- and intraobserver variability (*3*), principally arising from the poor quality of PET images. On the other hand, current state-of-the-art algorithms for functional-volume segmentation consist of fixed- (*12*) or adaptive-threshold approaches (*13,14*). Although fixed-threshold approaches are attractive because of their simplicity, their drawbacks are numerous given that the value of the threshold to be used for each lesion clearly depends on multiple factors, such as lesion contrast and size and image noise (*15*). The solutions based on the use of

adaptive thresholding consider the contrast between the object to delineate and its surrounding background. However, adaptive thresholding requires imaging system–specific optimization using uniformly filled spheric lesions, hence reducing the robustness of the approach, particularly in the case of multicenter trials. In addition, this method depends on the background ROI choice, which can in turn lead to reduced interobserver reproducibility for functional-volume determination. A few automatic algorithms have been proposed (*16–19*). The main difference between these algorithms and the threshold-based approaches is that the algorithms automatically estimate the parameters of interest and find the optimal regions' characteristics in a given image, without system-dependent parameters. This technique may reduce issues associated with deterministic approaches based on thresholding, potentially increasing the robustness and reproducibility of PET functional-volume determination (*20*).

Establishing the level of reproducibility and repeatability is essential in the use of any image-derived index in prognostic or therapy-response studies, allowing the evaluation of which change between 2 studies can be considered significant. To date, only a few reproducibility studies (*21–25*), almost exclusively concentrating on $SUV_{max}$ and $SUV_{mean}$ variability in double-baseline $^{18}$F-FDG PET scans, have shown a relative absolute percentage difference of up to 13%, with an SD of 10%. The reproducibility of quantitative indices (Patlak influx constant), associated with the acquisition of dynamic datasets, has also been assessed (*21,22*), showing similar levels of reproducibility (mean percentage difference, 8%–10%). Studies on the reproducibility of such indices in the case of $^{18}$F-FLT PET have shown that changes larger than 15%–20% and 25%–30% may be considered significant in $SUV_{mean}$ (obtained using a 41% fixed threshold) and $SUV_{max}$ or Patlak influx constant, respectively (*26,27*).

In most of these studies, $SUV_{mean}$ has been calculated using manually drawn ROIs or a single fixed threshold (varying from 40% to 75% of the maximum activity). Among these studies, only 1 has considered the reproducibility of metabolic functional volumes using a fixed threshold. Krak et al. (*3*) have shown a mean percentage difference in the ROI volumes of 23% ± 20% and 55% ± 35% for a fixed threshold of 50% and 75%, respectively. Finally, to our knowledge there has been no published study evaluating the reproducibility of TGV and TPV.

To date, despite numerous studies assessing the accuracy of different segmentation algorithms, there is a lack of evaluation of the repeatability and reproducibility of these algorithms relative to different threshold- and automatic-based delineation approaches. Therefore, the main objective of this study was to assess the repeatability and reproducibility in determining 3-dimensional (3D) functional volumes and associated indices ($SUV_{mean}$, TGV, and TPV) in PET using different algorithms. The reproducibility of $SUV_{max}$ was also included because it represents the index most used today in clinical practice and facilitates a direct comparison with previous studies. This evaluation was performed on double-baseline $^{18}$F-FDG and $^{18}$F-FLT clinical PET datasets.

## MATERIALS AND METHODS

### Segmentation Algorithms Considered

Four approaches were used in this work. Two different fixed thresholds (*12*) were considered, at 42% (T42) and 50% (T50) of the maximum voxel value, using a region-growing algorithm with the maximum-intensity voxel as seed.

An adaptive-threshold method (TSBR, for threshold source–to–background ratio) (*13*) was also included:

$$I_{threshold} = a + b\frac{1}{SBR}. \quad \text{Eq. 1}$$

SBR is the source-to-background ratio, defined as the contrast between a manually defined background ROI and the mean of the maximum-intensity voxel and its 8 surrounding neighbors in the same slice. The parameters a and b are optimized through linear regression analysis for a given scanner using phantom acquisitions of various sphere sizes and contrast.

For automatic-segmentation approaches, the fuzzy C-means (FCM) (*28*) clustering algorithm, with 2 clusters (background and lesion), was considered. This algorithm has been previously used for functional-volume segmentation tasks in both brain and oncology applications (*29,30*) and iteratively minimizes a cost function of the voxel-intensity values to estimate the center of each cluster and membership of each voxel to these clusters. The second automatic algorithm considered was the fuzzy locally adaptive Bayesian (FLAB) (*19*) methodology, based on a combination of statistical models with a fuzzy measure to simultaneously address issues of both noise and blur resulting from partial-volume effects in PET images. FLAB is also able to deal with strongly heterogeneous uptake in tumors of complex shape and generate nonbinary segmented volumes by considering 3 classes and the associated fuzzy transitions (*31*). The parameters required for the segmentation (gaussian mean and variance of each class and spatial priors for each voxel) were estimated using the iterative stochastic expectation maximization procedure. For all approaches, the tumors were delineated after having been isolated in a 3D box of interest previously defined and fixed for all segmentation methodologies (manual and automatic).

### Repeatability and Reproducibility: Definitions

Within the context of this study, repeatability is defined as the ability of a given segmentation algorithm to reach the same result regarding the definition of a functional volume when applied multiple times on a single image. In such a task, entirely deterministic fixed-threshold approaches (T42, T50) will always give the same result. On the other hand, more advanced methods—for example, the adaptive thresholding or automatic algorithms such as FCM and FLAB considered here—are susceptible to giving different results when applied multiple times on the same image. The adaptive-threshold segmentation, for instance, depends on a manually drawn background ROI and may thus result in variable delineation depending on the choice of this ROI. On the other hand, FCM and FLAB are iterative procedures that may not converge to the same result at each execution. Finally, manual delineation may be considered as the least repeat-

able, even when considering a single operator (intraoperator variability). A second aspect considered in this study was the impact of a segmentation algorithm on the reproducibility of determining functional volumes from 2 baseline PET scans.

Two different clinical datasets—comprising esophageal and breast cancer patients scanned with $^{18}$F-FDG and $^{18}$F-FLT, respectively—were used. In both cases, 2 consecutive PET scans were acquired at an interval of a few days. We therefore studied the differences in derived functional TVs, lesion $SUV_{mean}$, and TGVs and TPVs extracted from both images. The repeatability of measuring TVs using the various delineation approaches considered in this study was investigated for the same clinical datasets.

### Validation Studies

Fourteen whole-body $^{18}$F-FDG PET/CT images acquired for patients with esophageal cancer ($n$ = 17 lesions) and nine $^{18}$F-FLT PET/CT images acquired for breast cancer patients ($n$ = 12 lesions) were considered. Esophageal cancer patients' images were acquired at 3.4 ± 2.2 d on a PET/CT scanner (Gemini; Philips), with 2-min acquisitions per bed position, 60 min after the $^{18}$F-FDG injection (6 MBq/kg). Data were reconstructed using a 3D row-action maximization-likelihood algorithm with standard clinical protocol parameters (2 iterations, relaxation parameter of 0.05, 5 mm in full width at half maximum, 3D gaussian postfiltering). $^{18}$F-FLT PET images were acquired for patients with breast cancer (*27*); 2 scans were obtained within 2–7 d (median, 4.1 d) before treatment. All patients received a single bolus intravenous injection of $^{18}$F-FLT (153–381 MBq) over 30 s, and dynamic PET was performed for 95 min. Patients were scanned on a PET scanner (ECAT962/HR+; CTI/Siemens), and data were reconstructed using ordered-subset expectation maximization (360 iterations, 6 subsets, no postfiltering).

In both cases, 2 baseline scans were acquired within an average of 3–4 d of each other. Because no treatment was administered between the 2 baseline scans, and considering the short time between the 2 acquisitions, the assumption was that no significant physiologic changes occurred in between the time the scans were obtained. A similar assumption had been previously used in all other studies evaluating the reproducibility and repeatability of different SUV measurements in PET, with double-baseline scans obtained within 5–10 d (*21*–*25*). Figure 1 shows the 2 baseline scans—1 for an esophageal cancer (Fig. 1A) and 1 for a breast cancer (Fig. 1B) patient.

### Analysis

For the repeatability evaluation, the tumors in the first image for each patient were segmented 10 times each with FCM, FLAB, and TSBR. In addition, manual delineation was performed by 2 nuclear medicine experts. More specifically, the 2 experts performed 10 different slice-by-slice manual delineations for the different lesions considered in a randomized fashion, ensuring a minimum of a week between 2 consecutive delineations of the same lesion. All these manual segmentations were performed under the same conditions as those of full-range contrast display. The mean percentage variability and associated SD with respect to the mean segmented volume was computed for each of the lesions and segmentation approaches across the 10 executions and across the 10 manual delineations, to assess the repeatability of the approaches. The repeatability of the manual delineations of the 2 experts were compared separately (intraobserver variability) and with each other (interobserver variability) using intraclass coefficients.

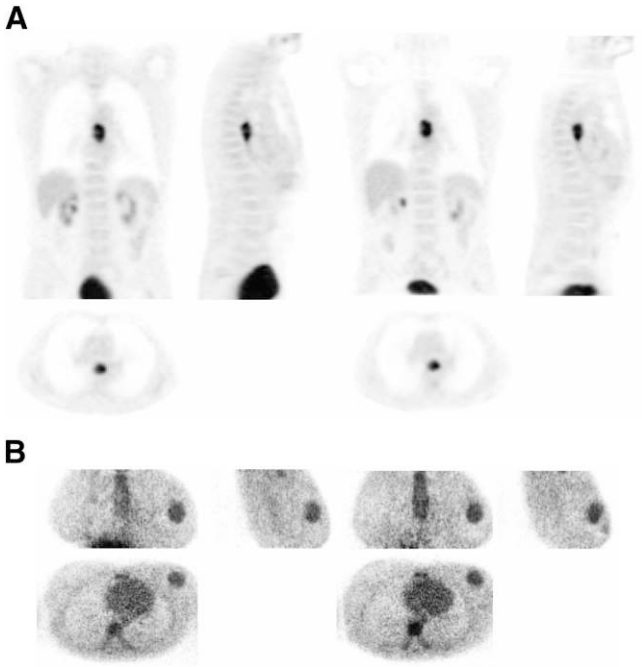



**FIGURE 1.** Baseline images: $^{18}$F-FDG (esophagus) (A) and $^{18}$F-FLT (breast) (B).

To study the relative impact of the different segmentation algorithms on the reproducibility of deriving different PET image indices, TVs were segmented independently on both baseline scan images for each lesion, using the different automatic-segmentation approaches. Subsequently, TV (in $cm^3$), $SUV_{mean}$, TGV or TPV, and $SUV_{max}$ quantitative values (M) were computed for each delineated lesion and compared between the 2 scans using the mean percentage difference relative to the mean of both baseline scans:

$$(M_{scan2} - M_{scan1}) \Big/ \frac{(M_{scan1} + M_{scan2})}{2} \times 100. \qquad \text{Eq. 2}$$

The distribution of the differences between each pair of measurements was assessed for each index using the Kolmogorov–Smirnov test, showing no significant differences from a normal distribution (Fig. 2). Bland–Altman analysis (*32*) was subsequently used to highlight differences between segmentation methodologies. Mean and SD of differences and the respective 95% confidence intervals (CIs) were obtained. To define the reproducibility limits (reference range of spontaneous changes), the 95% CIs for the difference between 2 measurements were computed as the mean difference ± 1.96 times the SD of the difference. To investigate any potential correlations in the measured reproducibility, the magnitude of the percentage difference for the TV, $SUV_{max}$, and $SUV_{mean}$ measurements was compared with the average of the TVs using the Pearson correlation coefficient $r$. This analysis was repeated to investigate the correlation of the reproducibility of the different parameters with the $SUV_{mean}$.

## RESULTS

Table 1 contains the mean variability and SD around the mean segmented volume across the 10 manual delineations

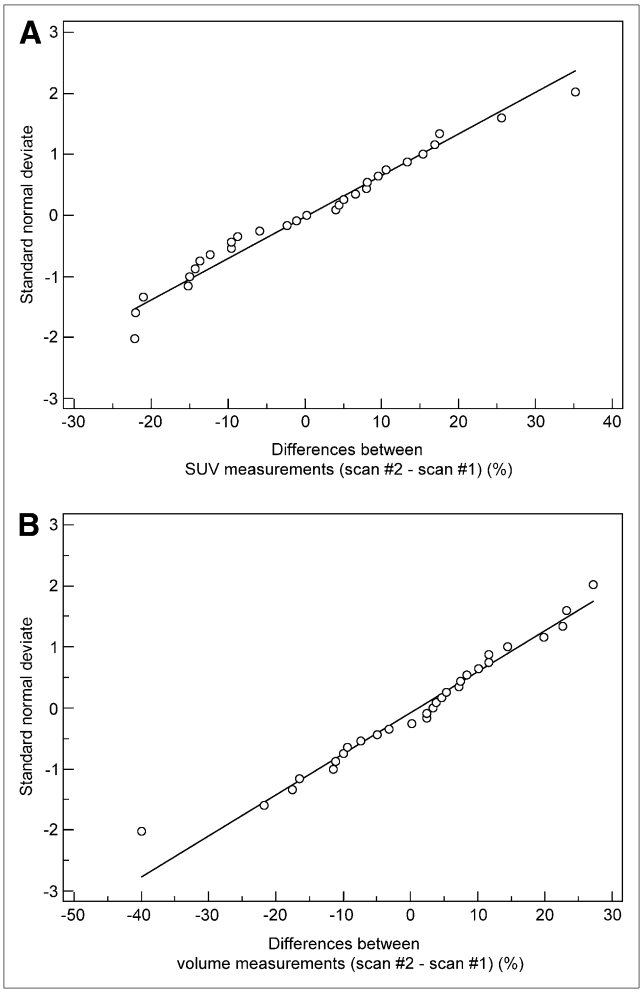



**FIGURE 2.** Plots showing that distributions of differences for $SUV_{mean}$ (FLAB) (A) and TV (FLAB) (B) between 2 scans were not significantly different from normal.

performed by each of the 2 nuclear medicine experts and 10 repeated executions of the FLAB, FCM, and TSBR algorithms. Results for both clinical datasets are presented separately. FLAB demonstrated highly repeatable results in all of the studied cases, with negligible variability (1%) around the mean segmented 3D volumes across the different repeated executions. FCM also led to satisfactory repeatability results (1.4% ± 1.6% for the $^{18}$F-FDG cases and 2.3% ± 1.9% for the $^{18}$F-FLT cases). In comparison, the use of the TSBR led to more than twice as high variability (2.9% ± 2.7% and 4.7% ± 3.6% for the $^{18}$F-FDG and $^{18}$F-FLT cases, respectively). By contrast, manual segmentation by the 2 experts showed high intraobserver variability for $^{18}$F-FDG esophageal lesions (14.1% ± 12.1% and 16.4% ± 11.3% for experts 1 and 2, respectively). Interobserver variability was 17.1% ± 14.3%, with an intraclass coefficient of 0.67 (95% CI, 0.39–0.89). In the case of $^{18}$F-FLT, this variability was even higher, with an intraobserver variability of 22.1% ± 18.7% and 23.8% ± 17.8% for experts 1 and 2, respectively, and an interobserver variability of 27.4% ± 21.9%, with an intraclass coefficient of 0.59 (95% CI, 0.31–0.84).

Tables 2 and 3 contain a summary of the reproducibility results for the different parameters computed from Bland–Altman plots on the 2 consecutive baseline scans for $^{18}$F-FDG esophageal and $^{18}$F-FLT breast lesions, respectively. The observed reproducibility of $SUV_{max}$ and $SUV_{mean}$ measurements for the volumes obtained using TSBR and FLAB is illustrated in Figure 3. The corresponding plots for TV are shown in Figures 4A and 4B using TSBR and FLAB, respectively.

Concerning the reproducibility of $SUV_{max}$, similar percentage differences were measured for the $^{18}$F-FDG and $^{18}$F-FLT datasets, with an SD of the mean percentage difference of 16.7% and 14.9%, respectively. The upper and lower percentage reproducibility limits for the $SUV_{max}$ were −31% to 35% and −30% to 28% for the $^{18}$F-FDG and $^{18}$F-FLT datasets, respectively. On the other hand, the automatic approaches led to $^{18}$F-FDG TV measurement reproducibility limits of −21% to 31% and −51% to 52% for the FLAB and the FCM algorithms, respectively. A poorer reproducibility of the $^{18}$F-FDG TV measurements was observed for the threshold-based approaches, with upper and lower reproducibility limits of −90% to 51% and −69% to 73% for the adaptive and T42, respectively. In the case of $^{18}$F-FLT TV measurements, the reproducibility was similar to that of $^{18}$F-FDG for the threshold-based approaches, whereas a deterioration in the reproducibility obtained with the automatic approaches was observed, particularly for the FCM algorithm (with reproducibility limits of −66% to 74%).

$SUV_{mean}$ measurements using FLAB exhibited reproducibility levels similar in magnitude to that for the TV definition, with an SD of the mean percentage difference of 15.6% and 14.1% for the $^{18}$F-FDG and $^{18}$F-FLT datasets, respectively. This was, however, not the case for the other tumor-delineation algorithms considered, with the larger $SUV_{mean}$ reproducibility limits using the FCM tumor definition (−77% to 62% and −59% to 59% for the $^{18}$F-FDG and $^{18}$F-FLT datasets, respectively). Finally, the smaller $SUV_{mean}$ reproducibility for the threshold-based approaches was obtained using T50 for both the $^{18}$F-FDG and the $^{18}$F-FLT datasets, with a mean percentage difference of −10.5% ± 23% and −13.3% ± 16.8%, respectively.

The reproducibility of TGV and TPV, being the product of TV and $SUV_{mean}$, was dependent on the direction of changes for both TV and $SUV_{mean}$. As an increase of TV was correlated with a decrease of $SUV_{mean}$ and vice versa ($P < 0.002$; $r = 0.54$, 0.67, and 0.72 for FLAB, TSBR, and T42, respectively), TGV and TPV reproducibility levels were generally similar in magnitude to the TV and $SUV_{mean}$ considered separately. However, in certain cases there were more increases or decreases of both TV and $SUV_{mean}$ for a given patient, resulting in larger variability of the TGV and TPV measurements (e.g., the TSBR measurements of the

**TABLE 1.** Repeatability Evaluation

| Method | Esophageal lesion | | Breast lesion | |
|---|---|---|---|---|
| | Mean variability (%) | SD | Mean variability (%) | SD |
| FLAB | 0.6 | 0.3 | 1.1 | 0.7 |
| FCM | 1.4 | 1.6 | 2.3 | 1.9 |
| Fixed threshold | 0 | 0 | 0 | 0 |
| Adaptive threshold | 2.9 | 2.7 | 4.7 | 3.6 |
| Manual delineation (expert 1) | 14.1 | 12.2 | 22.1 | 18.7 |
| Manual delineation (expert 2) | 16.4 | 11.3 | 23.8 | 17.8 |
| Manual delineation (expert 2 with respect to 1) | 17.1 | 14.3 | 27.4 | 21.9 |

Data are mean variability and SD around mean segmented volume for repeated delineations of 17 esophageal and 12 breast lesions on first baseline $^{18}$F-FDG and $^{18}$F-FLT scans, respectively.

$^{18}$F-FLT breast lesions, with 22.1% ± 48.9% for the TPV, whereas TV and $SUV_{mean}$ were 11.3% ± 31.4% and −3.2% ± 26.5%, respectively).

The TV reproducibility results were dependent on the measured TV, with a larger variability seen for smaller tumors. This dependence was statistically significant for the adaptive thresholding ($r = 0.37$, $P = 0.046$; Fig. 5A), with differences higher than 30% on average (≤75%) in several of the tumors below 50 cm$^3$. On the other hand, this dependence was not significant for FLAB ($r = 0.27$, $P = 0.16$; Fig. 5B), with most differences less than 30%—irrespective of TV—further demonstrating improved robustness, as previously shown (*19,20*). In terms of the $SUV_{max}$ reproducibility results, no statistically significant trend with either the lesion size ($r = 0.016$, $P = 0.93$; Fig. 5C) or the mean of the 2 $SUV_{mean}$ measurements ($r = 0.14$, $P = 0.49$) was observed. Finally, no statistically significant trends were found for the $SUV_{mean}$ reproducibility depending on the lesion size, irrespective of the segmentation algorithm used ($r = 0.2$, $P = 0.3$, and $r = 0.23$, $P = 0.23$, for TSBR and FLAB, respectively).

## DISCUSSION

Functional-volume delineation today represents an area of interest for multiple clinical (routine and research) applications of PET (prognosis, response prediction, therapy assessment, radiotherapy treatment planning). In all of these applications, the repeatability and reproducibility with which functional volumes can be determined under different imaging conditions play a predominant role, allowing a level of confidence to be established in the use of such TV measurements. Volume-definition methodologies currently used in clinical practice are based on the use of manual delineation or fixed and adaptive thresholding (*12–14*), whereas several promising automatic algorithms have been proposed (*16–19*). The major drawback of manual delineation is high inter- and intraobserver variability; in addition, the approach is time-consuming. On the other

**TABLE 2.** Reproducibility Results Using $^{18}$F-FDG for Esophageal Lesions

| Method | Parameter | Mean ± SD | 95% CI | LRL | 95% CI for LRL | URL | 95% CI for URL |
|---|---|---|---|---|---|---|---|
| | $SUV_{max}$ | 1.8 ± 16.7 | −6.8 to 10.4 | −30.9 | −45.9 to −16 | 34.6 | 19.9–49.6 |
| FLAB | TV | 5 ± 13.3 | −1.8 to 11.9 | −21.1 | −33 to −9.1 | 31.1 | 19.2–43 |
| | $SUV_{mean}$ | 0 ± 15.6 | −8 to 8 | −30.5 | −44.4 to −16.6 | 30.5 | 16.5–44.4 |
| | TGV | 5.1 ± 10.6 | −0.4 to 10.5 | −15.8 | −25.3 to −6.3 | 25.9 | 16.4–35.5 |
| FCM | TV | 0.4 ± 26.4 | −13.2 to 14 | −51.4 | −75.1 to −27.7 | 52.2 | 28.5–75.9 |
| | $SUV_{mean}$ | −7.8 ± 35.5 | −26 to 10.5 | −77.4 | −109.2 to −45.5 | 61.8 | 30–93.7 |
| | TGV | −7.4 ± 30.2 | −22.9 to 8.2 | −66.6 | −93.7 to −39.5 | 51.9 | 24.8–78.9 |
| TSBR | TV | −19.4 ± 36 | −37.9 to −0.9 | −89.9 | −122.1 to −57.6 | 51.1 | 18.9–83.3 |
| | $SUV_{mean}$ | 6.3 ± 27.4 | −7.8 to 20.4 | −47.4 | −72 to −22.8 | 60.1 | 35.5–84.6 |
| | TGV | −13 ± 28.2 | −27.5 to 1.5 | −68.2 | −93.4 to −42.9 | 42.2 | 17–67.4 |
| T42 | TV | 2.1 ± 36.1 | −16.5 to 20.7 | −68.7 | −101.2 to −36.3 | 72.9 | 40.5–105.3 |
| | $SUV_{mean}$ | −10.5 ± 30 | −25.9 to 5 | −69.3 | −96.2 to −42.4 | 48.4 | 21.5–75.3 |
| | TGV | −8.4 ± 23.4 | −20.5 to 3.6 | −54.3 | −75.3 to −33.3 | 37.5 | 16.5–58.5 |
| T50 | TV | 0.9 ± 32.9 | −16 to 17.8 | −63.5 | −92.9 to −34 | 65.3 | 35.9–94.8 |
| | $SUV_{mean}$ | −10.5 ± 23 | −22.6 to 1.6 | −56.5 | −77.6 to −35.5 | 35.6 | 14.5–56.6 |
| | TGV | −9.5 ± 23.1 | −21.4 to 2.4 | −54.9 | −75.6 to 34.1 | 35.8 | 15.1–56.6 |

LRL = lower reproducibility limit; URL = upper reproducibility limit.
Data are percentage differences between scan 2 and scan 1 measurements.

**TABLE 3.** Reproducibility Results Using $^{18}$F-FLT for Breast Lesions

| Method | Parameter | Mean ± SD | 95% CI | LRL | 95% CI for LRL | URL | 95% CI for URL |
|---|---|---|---|---|---|---|---|
| | $SUV_{max}$ | −0.9 ± 14.9 | −10.4 to 8.5 | −30 | −46.6 to −13.4 | 28.2 | 11.6–44.8 |
| FLAB | TV | 4.3 ± 15.7 | −5.7 to 14.3 | −26.5 | −44.1 to −8.9 | 35.2 | 17.6–52.8 |
| | $SUV_{mean}$ | −0.6 ± 14.1 | −9.6 to 8.3 | −28.2 | −44 to −12.5 | 27 | 11.2–42.7 |
| | TGV | 3.7 ± 17.2 | −7.2 to 14.6 | −30 | −49.2 to −10.8 | 37.4 | 18.2–56.6 |
| FCM | TV | 4.2 ± 35.7 | −18.4 to 26.9 | −65.6 | −105.5 to −25.8 | 74.1 | 34.3–114 |
| | $SUV_{mean}$ | 0.3 ± 30.1 | −18.8 to 19.4 | −58.6 | −92.2 to −25 | 59.2 | 25.6–92.8 |
| | TGV | 4.6 ± 29.8 | −14.3 to 23.6 | −53.9 | −87.2 to −20.5 | 63.1 | 29.7–96.4 |
| TSBR | TV | 11.3 ± 31.4 | −8.7 to 31.2 | −50.4 | −85.5 to −15.2 | 72.8 | 37.7–108 |
| | $SUV_{mean}$ | −3.2 ± 26.5 | −20 to 16.6 | −55.1 | −84.7 to −25.5 | 48.7 | 19.1–78.3 |
| | TGV | 22.1 ± 48.9 | −9 to 53.2 | −73.8 | −128.5 to −19.1 | 118 | 63.3–172.7 |
| T42 | TV | 9.8 ± 35 | −12.4 to 32.1 | −58.7 | −97.8 to −19.6 | 78.4 | 39.3–117.5 |
| | $SUV_{mean}$ | −9.4 ± 20.9 | −22.7 to 3.9 | −50.3 | −73.7 to −27 | 31.6 | 8.2–54.9 |
| | TGV | 0.7 ± 27.3 | −16.7 to 18 | −52.8 | −83.3 to −22.3 | 54.1 | 23.6–84.6 |
| T50 | TV | 11.2 ± 31.4 | −8.8 to 31.1 | −50.5 | −85.6 to −15.3 | 72.8 | 37.6–107.9 |
| | $SUV_{mean}$ | −13 ± 16.8 | −24 to −2.7 | −46.2 | −64.9 to −27.4 | 19.5 | 0.8–38.3 |
| | TGV | −1.8 ± 26 | −18.4 to 14.7 | −52.8 | −81.9 to −23.7 | 49.1 | 20.1–78.2 |

LRL = lower reproducibility limit; URL = upper reproducibility limit.
Data are percentage differences between scan 2 and scan 1 measurements.

hand, currently considered state-of-the art adaptive threshold–based algorithms have been shown to accurately define functional volumes under certain imaging conditions of spheric and homogeneous-activity-distribution lesions. However, adaptive-threshold approaches usually involve some user interaction to select background ROIs, which can potentially lead to user-introduced variability. Although signal intensity reproducibility, predominantly considering the use of $SUV_{max}$, has been previously assessed, the potential of new indices such as TV or TGV and TPV can be considered only after the assessment of their reproducibility, which has not been previously widely assessed. Therefore, in this study the reproducibility limits of these indices, in comparison to other indices considered as the current gold standard, have been assessed using different tumor-delineation methodologies on double-baseline $^{18}$F-FDG and $^{18}$F-FLT datasets.

In terms of repeatability, all algorithms exhibited mean differences of less than 5%, with automatic approaches coming closer to the perfect repeatability that can be achieved by deterministic approaches such as a fixed threshold. The repeatability of both threshold and automatic-segmentation approaches was superior to that of manual delineation. This should, of course, be considered within the context of the limited absolute accuracy of thresholding, particularly for lesions not homogeneous in form and activity distribution (*31*).

The variability in the $SUV_{max}$ observed in this work is similar to that measured in previous reproducibility studies, with comparable percentage differences for $^{18}$F-FDG and $^{18}$F-FLT datasets. These percentage differences suggest that differences larger than −30% can be considered as significant in treatment response, whereas changes above 35% (30% for $^{18}$F-FLT) may be indicative of no response. Depending on the delineation algorithm used, the mean percentage difference and corresponding SD for TV measured on the 2 baseline scans varied from 5% ± 13% to −19% ± 36% for the $^{18}$F-FDG and from 4% ± 16% to 10% ± 35% for the $^{18}$F-FLT datasets. The smallest TV reproducibility limits obtained were similar to those for $SUV_{max}$. These limits ranged from −21% to 31% and −27% to 35% for $^{18}$F-FDG and $^{18}$F-FLT, respectively, suggesting in turn that, depending on the segmentation algorithm used and similar to $SUV_{max}$, CIs may be considered for monitoring therapy response based on functional TV. Similarly, in the case of TGV and TPV the smallest reproducibility limits measured were between −16% to 26% and −30% to 37% for $^{18}$F-FDG and $^{18}$F-FLT, respectively. On the other hand, the largest reproducibility limits for the $^{18}$F-FDG TV and TGV ranged from −90% to 73% and from −68% to 52%, respectively.

Reproducibility ranges obtained for the $^{18}$F-FDG esophageal lesions were almost systematically smaller than the ones obtained on the $^{18}$F-FLT breast lesions—which can be attributed to the higher level of noise and overall lower contrast observed in the $^{18}$F-FLT cases, resulting in less robust delineations. In addition, $^{18}$F-FDG esophageal lesions tended to appear more homogeneous than breast lesions. For instance, FCM—which incorporates neither noise nor spatial modeling—is associated with a larger mean TV variability of the $^{18}$F-FLT dataset relative to $^{18}$F-FDG, whereas FLAB exhibited similar reproducibility levels for both. The variability in reproducibility highlights the need for a robust delineation tool ensuring high reproducibility in an environment of substantial image-quality variability—likely, for example, to be encountered in multicenter trials in which the use of functional TV as a measure of response to therapy may be considered.

T50 uses a more restrictive threshold than 42% and is therefore less prone to large overevaluation of low contrast

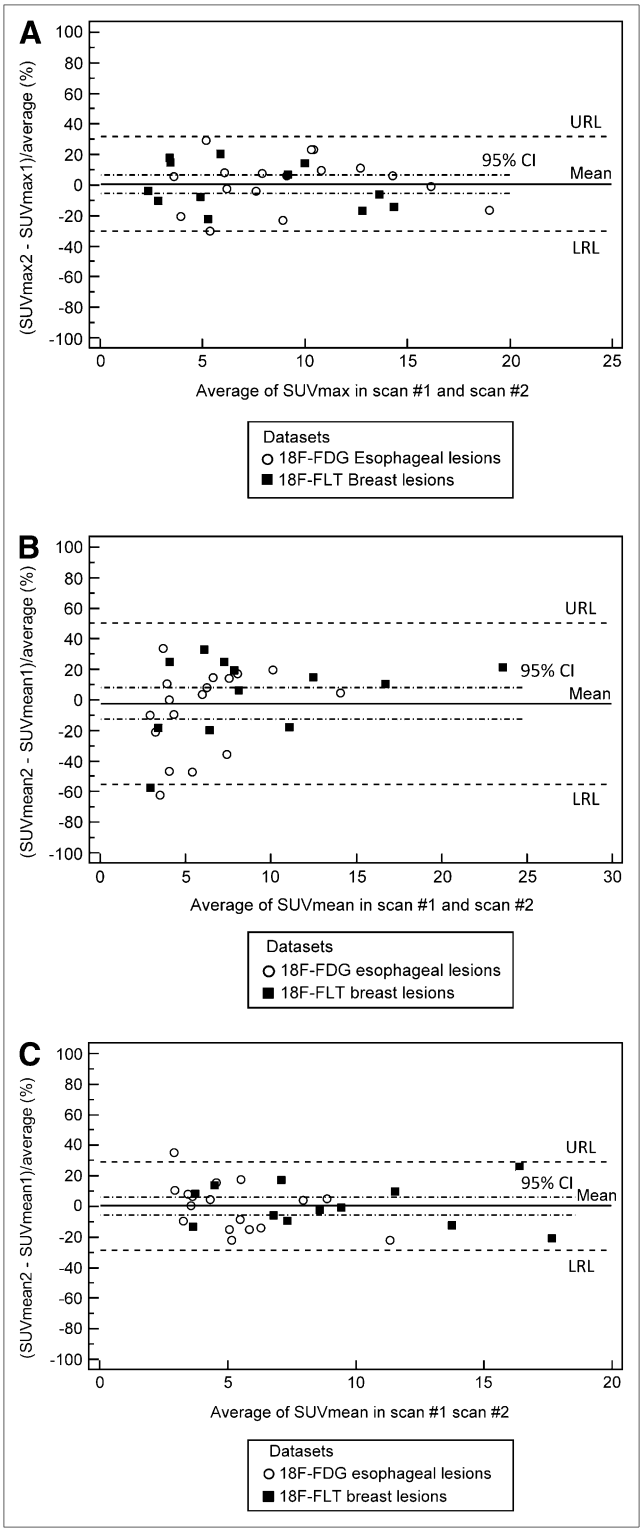



**FIGURE 3.** Bland–Altman plots of $SUV_{max}$ (A), $SUV_{mean}$ using adaptive thresholding (B), and $SUV_{mean}$ using FLAB (C) for both $^{18}$F-FDG and $^{18}$F-FLT lesions. Lines show combined mean, 95% CI, and upper and lower reproducibility limits. Individual values for $^{18}$F-FDG and $^{18}$F-FLT lesions are shown in Tables 2 and 3, respectively. LRL = lower reproducibility limit; URL = upper reproducibility limit.

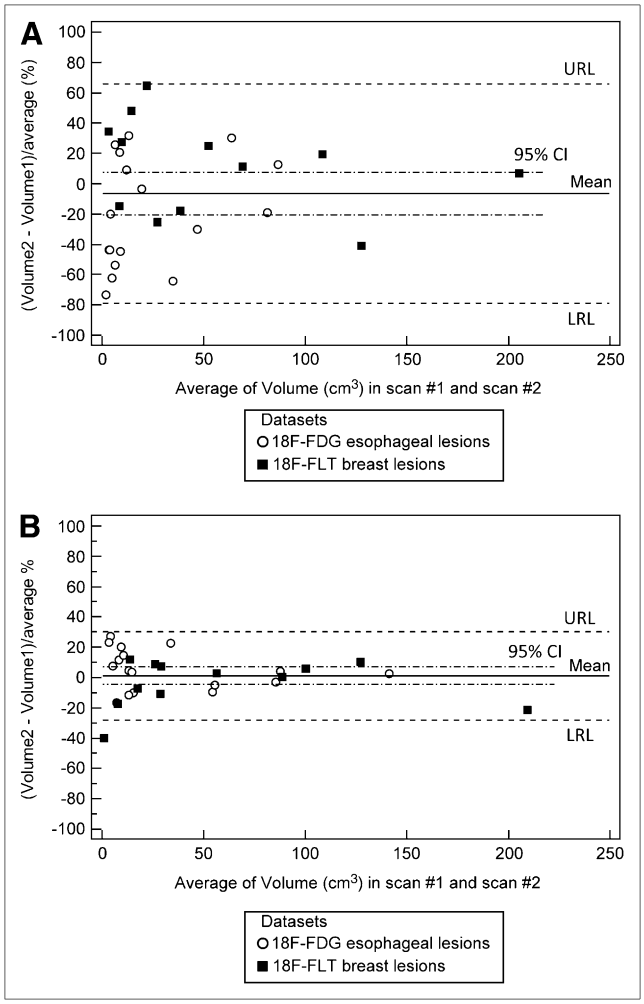



**FIGURE 4.** Bland–Altman plots of TV using adaptive thresholding (A) and TV using FLAB (B) for both $^{18}$F-FDG and $^{18}$F-FLT lesions. Lines show combined mean, 95% CI, and upper and lower reproducibility limits. Individual values for $^{18}$F-FDG and $^{18}$F-FLT lesions are shown in Tables 2 and 3, respectively. LRL = lower reproducibility limit; URL = upper reproducibility limit.

(<4:1) or small-size (<2 cm in diameter) TVs. T50 led to systematically lower variability than T42. Finally, the adaptive-threshold methodology did not demonstrate better reproducibility than did fixed thresholding, which can be attributed to the use of the background ROI placed manually on both scans, combined with the fact that background activity may also vary between the 2 scans.

Although a potential criticism for the current study can be the lack of ground-truth for the functional volumes, the aim of this work was not to assess the absolute accuracy of algorithms, which has been assessed previously for the approaches used in this work (*19,31*). The objective was to assess the reproducibility limits of functional-volume–related indices that can be attained depending on the algorithm. Within this context, the repeated studies of the double-baseline acquisitions have been performed within an

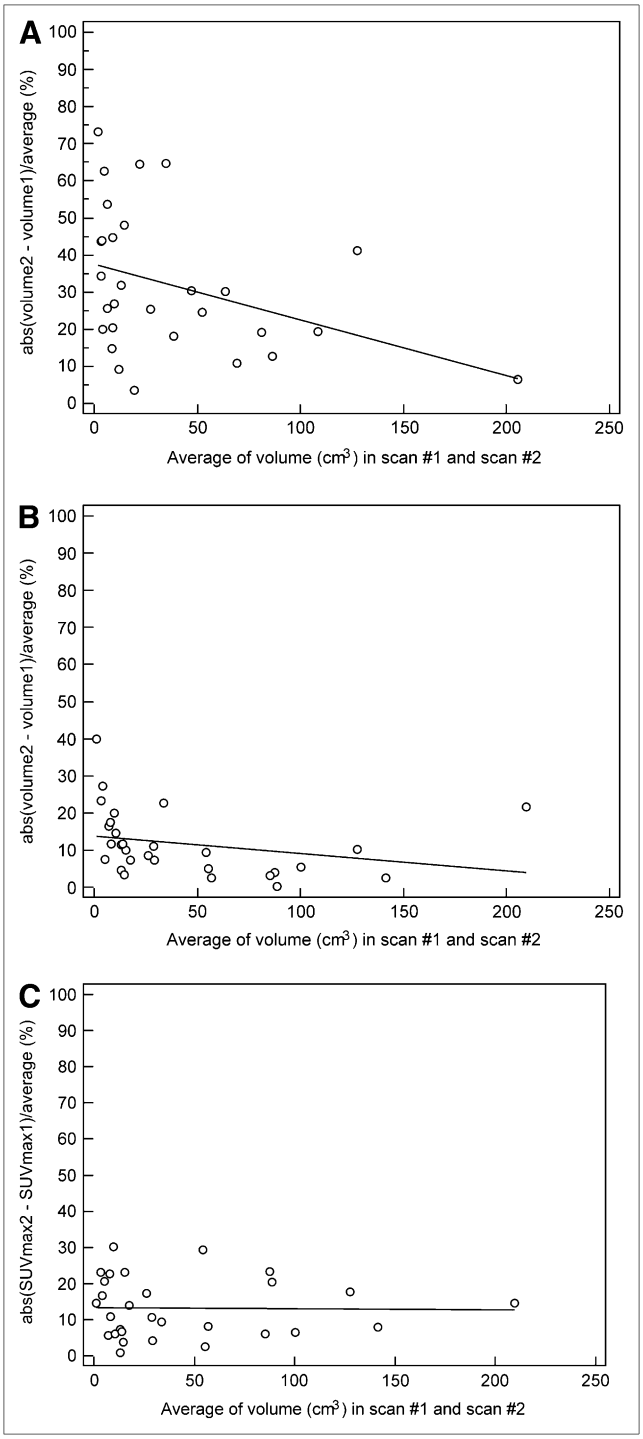



**FIGURE 5.** Differences between TVs (A and B) and $SUV_{max}$ (C) measured in 2 baseline scans in relation to average TV obtained using adaptive thresholding (A) and FLAB (B and C). abs = absolute.

average of 3–4 d, without any treatment between them, matching the method used by all other reproducibility studies to date (*21–25*). Finally, the reproducibility of $SUV_{max}$ was included in this work as the current gold standard, facilitating at the same time the comparison of our reproducibility study to those performed previously. The $SUV_{max}$ reproducibility limits obtained in this work for both $^{18}$F-FDG and $^{18}$F-FLT agree closely with those of previous studies.

## CONCLUSION

The smaller reproducibility ranges obtained for the different image indices considered in this study, similar to those of $SUV_{max}$, suggest that new automatic-segmentation approaches may facilitate the introduction of TVs or a combination of TVs and signal intensity in the form of TGVs and TPVs derived from PET images for therapy-response studies. However, our results also demonstrate that the reproducibility of different quantitative parameters associated with functional volumes depends significantly on the delineation approach.

## ACKNOWLEDGMENTS

We gratefully acknowledge funding by the Ligue Contre le Cancer (Finistère Committee), French National Research Agency (ANR-08-ETEC-005-01), Cancéropôle Grand Ouest (R05014NG), CR-UK & EPSRC Cancer Imaging Centre (Imperial College, London), U.K. Medical Research Council, and Department of Health (C2536/A10337, U.1200.02.005.00001.01).

## REFERENCES

1. Kelloff GJ, Hoffman JM, Johnson B, et al. Progress and promise of FDG PET imaging for cancer patient management and oncologic drug development. *Clin Cancer Res.* 2005;11:2785–2808.
2. Visvikis D, Cheze-Le Rest C, Costa DC, Bomanji J, Gacinovic S, Ell PJ. Influence of OSEM and segmented attenuation correction in the calculation of standardised uptake values for $^{18}$FDG-PET. *Eur J Nucl Med Mol Imaging.* 2001; 28:1326–1335.
3. Krak NC, Boellaard R, Hoekstra OS, Twisk JW, Hoekstra CJ, Lammertsma AA. Effects of ROI definition and reconstruction method on quantitative outcome and applicability in a response monitoring trial. *Eur J Nucl Med Mol Imaging.* 2005; 32:294–301.
4. Lucignani G, Larson SM. Doctor, what does my future hold? The prognostic values of FDG-PET in solid tumours. *Eur J Nucl Med Mol Imaging.* 2010;37: 1032–1038.
5. Wahl RL, Jacene H, Kasamon Y, Lodge MA. From RECIST to PERCIST: evolving considerations for PET response criteria in solid tumors. *J Nucl Med.* 2009;50(suppl 1):122S–150S.
6. Seol YM, Kwon BR, Song MK, et al. Measurement of tumor volume by PET to evaluate prognosis in patients with head and neck cancer treated by chemo-radiation therapy. *Acta Oncol.* 2010;49:201–208.
7. Chung MK, Jeong HS, Park SG, et al. Metabolic tumor volume of [$^{18}$F]-fluorodeoxyglucose positron emission tomography/computed tomography predicts short-term outcome to radiotherapy with or without chemotherapy in pharyngeal cancer. *Clin Cancer Res.* 2009;15:5861–5868.
8. Hyun SH, Choi JY, Shim YM, et al. Prognostic value of metabolic tumor volume measured by $^{18}$F-fluorodeoxyglucose positron emission tomography in patients with esophageal carcinoma. *Ann Surg Oncol.* 2010;17:115–122.
9. Larson SM, Erdi Y, Akhurst T, et al. Tumor treatment response based on visual and quantitative changes in global tumor glycolysis using PET FDG imaging: the visual response score and the change in total lesion glycolysis. *Clin Positron Imaging.* 1999;2:159–171.
10. Francis RJ, Byrne MJ, Van der Schaaf AA, et al. Early prediction of response to chemotherapy and survival in malignant pleural mesothelioma using a novel semiautomated 3-dimensional volume-based analysis of serial $^{18}$F-FDG PET scans. *J Nucl Med.* 2007;48:1449–1458.
11. Cazaentre T, Morschhauser F, Vermandel M, et al. Pre-therapy $^{18}$F-FDG PET quantitative parameters help in predicting the response to radioimmunotherapy in non-Hodgkin lymphoma. *Eur J Nucl Med Mol Imaging.* 2010;37:494–504.

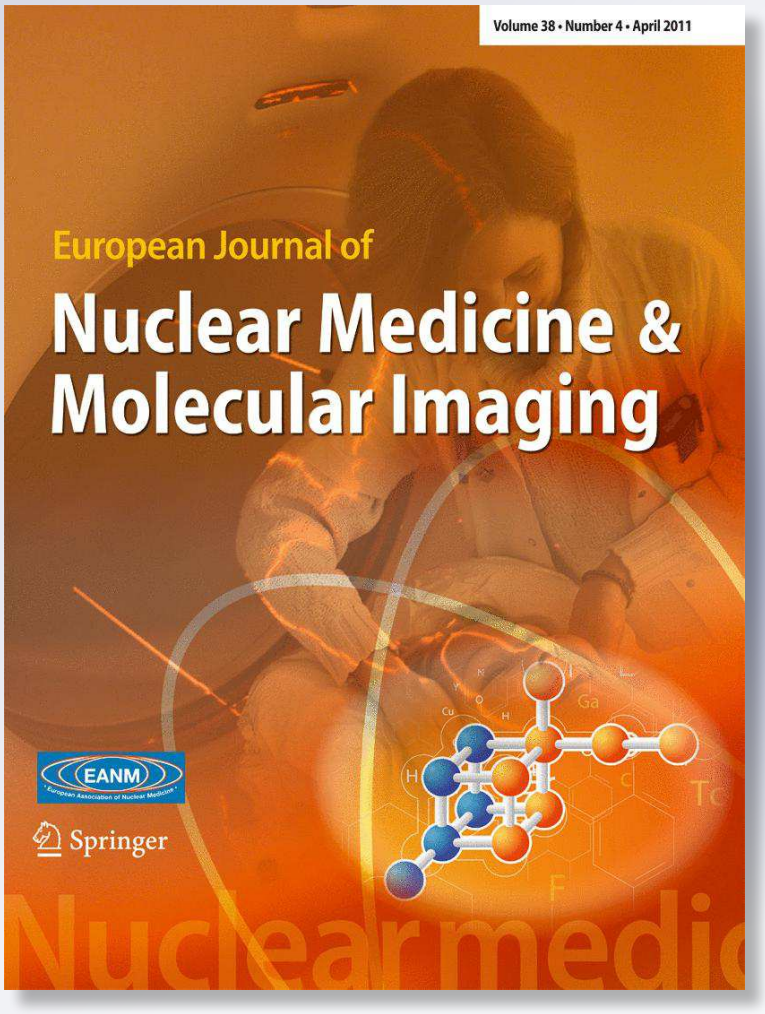
12. Erdi YE, Mawlawi O, Larson SM, et al. Segmentation of lung lesion volume by adaptive positron emission tomography image thresholding. *Cancer.* 1997;80 (suppl 12):2505–2509.
13. Daisne J-F, Sibomana M, Bol A, et al. Tri-dimensional automatic segmentation of PET volumes based on measured source-to-background ratios: influence of reconstruction algorithms. *Radiother Oncol.* 2003;69:247–250.
14. Nestle U, Kremp S, Schaefer-Schuler A, et al. Comparison of different methods for delineation of $^{18}$F-FDG PET-positive tissue for target volume definition in radiotherapy of patients with non-small cell lung cancer. *J Nucl Med.* 2005;46: 1342–1348.
15. Biehl KJ, Kong FM, Dehdashti F, et al. $^{18}$F-FDG PET definition of gross tumor volume for radiotherapy of non-small cell lung cancer: is a single standardized uptake value threshold approach appropriate? *J Nucl Med.* 2006;47:1808–1812.
16. El Naqa I, Yang D, Apte A, et al. Concurrent multimodality image segmentation by active contours for radiotherapy treatment planning. *Med Phys.* 2007;34: 4738–4749.
17. Montgomery DWG, Amira A, Zaidi H. Fully automated segmentation of oncological PET volumes using a combined multiscale and statistical model. *Med Phys.* 2007;34:722–736.
18. Geets X, Lee JA, Bol A, et al. A gradient-based method for segmenting FDG-PET images: methodology and validation. *Eur J Nucl Med Mol Imaging.* 2007;34:1427–1438.
19. Hatt M, Turzo A, Roux C, et al. A fuzzy Bayesian locally adaptive segmentation approach for volume determination in PET. *IEEE Trans Med Imaging.* 2009;28:881–893.
20. Hatt M, Bailly P, Turzo A, Roux C, Visvikis D. Automatic delineation of functional volumes in PET: a robustness study [abstract]. *J Nucl Med.* 2009;50 (suppl 2):282P.
21. Minn H, Clavo AC, Grenman R, Wahl RL. In vitro comparison of cell proliferation kinetics and uptake of tritiated fluorodeoxyglucose and L-methionine in squamous-cell carcinoma of the head and neck. *J Nucl Med.* 1995;36:252–258.
22. Weber WA, Ziegler SI, Thodtmann R, Hanauske AR, Schwaiger M. Reproducibility of metabolic measurements in malignant tumors using FDG PET. *J Nucl Med.* 1999;40:1771–1777.
23. Nahmias C, Wahl LM. Reproducibility of standardized uptake value measurements determined by $^{18}$F-FDG PET in malignant tumors. *J Nucl Med.* 2008;49:1804–1808.
24. Paquet N, Albert A, Foidart J, Hustinx R. Within patient variability of FDG standardized uptake values in normal tissues. *J Nucl Med.* 2004;45:784–788.
25. Velasquez LM, Boellaard R, Kollia G, et al. Repeatability of $^{18}$F-FDG PET in a multicenter phase I study of patients with advanced gastrointestinal malignancies. *J Nucl Med.* 2009;50:1646–1654.
26. De Langen AJ, Klabbers B, Lubberink M, et al. Reproducibility of quantitative 18FLT measurements using positron emission tomography. *Eur J Nucl Med Mol Imaging.* 2009;36:389–395.
27. Kenny L, Coombes RC, Vigushin DM, et al. Imaging early changes in proliferation at 1 week post chemotherapy: a pilot study in breast cancer patients with FLT positron emission tomography. *Eur J Nucl Med Mol Imaging.* 2007;34:1339–1347.
28. Dunn JC. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *J Cybernet.* 1973;31:32–57.
29. Zhu W, Jiang T. Automation segmentation of PET image for brain tumors. *IEEE Nucl Sci Symp Conf Rec.* 2003;4:2627–2629.
30. Belhassen S, Zaidi H. Segmentation of heterogeneous tumors in PET using a novel fuzzy C-means algorithm [abstract]. *J Nucl Med.* 2009;50(suppl 2):286P.
31. Hatt M, Cheze-le Rest C, Descourt P, et al. Accurate automatic delineation of heterogeneous functional volumes in positron emission tomography for oncology applications. *Int J Radiat Oncol Biol Phys.* 2010;77:301–308.
32. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet.* 1986;1:307–310.

# *PET functional volume delineation: a robustness and repeatability study*

**European Journal of Nuclear Medicine and Molecular Imaging**

ISSN 1619-7070
Volume 38
Number 4

Eur J Nucl Med Mol Imaging
(2011) 38:663-672
DOI 10.1007/
s00259-010-1688-6




Springer

Eur J Nucl Med Mol Imaging (2011) 38:663–672
DOI 10.1007/s00259-010-1688-6

ORIGINAL ARTICLE

# PET functional volume delineation: a robustness and repeatability study

**Mathieu Hatt · Catherine Cheze Le Rest · Nidal Albarghach · Olivier Pradier · Dimitris Visvikis**

Received: 17 September 2010 / Accepted: 15 November 2010 / Published online: 12 January 2011


**Abstract**
*Purpose* Current state-of-the-art algorithms for functional uptake volume segmentation in PET imaging consist of threshold-based approaches, whose parameters often require specific optimization for a given scanner and associated reconstruction algorithms. Different advanced image segmentation approaches previously proposed and extensively validated, such as among others fuzzy C-means (FCM) clustering, or fuzzy locally adaptive bayesian (FLAB) algorithm have the potential to improve the robustness of functional uptake volume measurements. The objective of this study was to investigate robustness and repeatability with respect to various scanner models, reconstruction algorithms and acquisition conditions.
*Methods and materials* Robustness was evaluated using a series of IEC phantom acquisitions carried out on different PET/CT scanners (Philips Gemini and Gemini Time-of-Flight, Siemens Biograph and GE Discovery LS) with their associated reconstruction algorithms (RAMLA, TF MLEM, OSEM). A range of acquisition parameters (contrast, duration) and reconstruction parameters (voxel size) were considered for each scanner model, and the repeatability of each method was evaluated on simulated and clinical tumours and compared to manual delineation.
*Results* For all the scanner models, acquisition parameters and reconstruction algorithms considered, the FLAB algorithm demonstrated higher robustness in delineation of the spheres with low mean errors (10%) and variability (5%), with respect to threshold-based methodologies and FCM. The repeatability provided by all segmentation algorithms considered was very high with a negligible variability of <5% in comparison to that associated with manual delineation (5–35%).
*Conclusion* The use of advanced image segmentation algorithms may not only allow high accuracy as previously demonstrated, but also provide a robust and repeatable tool to aid physicians as an initial guess in determining functional volumes in PET.

**Keywords** PET uptake volume determination · Robustness · Repeatability · FLAB · Thresholding

M. Hatt · C. Cheze Le Rest · N. Albarghach · O. Pradier · D. Visvikis
INSERM, U650, LaTIM, CHU Morvan,
Brest 29200, France

C. Cheze Le Rest
Academic Department of Nuclear Medicine, CHU,
Brest 29200, France

N. Albarghach · O. Pradier
Institute of Oncology, CHU,
Brest 29200, France

M. Hatt (✉)
LaTIM, INSERM U650, CHU MORVAN,
5 avenue Foch,
29609 Brest, France
e-mail: hatt@univ-brest.fr

## Introduction

Accurate, robust, reproducible and fast delineation of functional tumour uptake volumes in three dimensions using positron emission tomography (PET) has been identified as a pressing challenge for an increasing number of oncology applications, such as image-guided radiotherapy [1–3], diagnosis, prognosis and therapy response assessment [4, 5]. On the one hand, manual delineation of functional uptake volumes using PET images is tedious and associated with very low repeatability due to high inter- and

intraobserver variability [4], principally arising from the poor quality of PET images. On the other hand, current state-of-the-art algorithms for functional uptake volume segmentation using PET images consist of fixed [6] or adaptive thresholding approaches [7, 8]. Regarding the use of a fixed threshold, numerous studies have shown the need for a variable threshold, depending on numerous factors, such as among others, lesion contrast, lesion size, and image noise [9]. As a solution, in the case of adaptive thresholding, the applied threshold depends on the measured contrast between the object delineated and its surrounding background, as well as parameters requiring optimization on phantom acquisitions. This optimization has to be performed for each scanner model and associated reconstruction and correction algorithms, making these approaches system-dependent. In addition, recent studies have shown that even for the same scanner model, a significant variation in the "ideal" threshold may exist due to differences in clinical acquisition and reconstruction protocols [10] underlining the possibility that such deterministic approaches may not be sufficiently robust and reproducible for functional uptake volume determination.

Recently several advanced image segmentation algorithms have been proposed in the literature for PET volume delineation [11–16]. The physical accuracy of these algorithms in differentiating the uptake signal from its surrounding background has, in most cases, already been assessed with respect to ground-truth, provided by a combination of realistic simulated or acquired phantom images as well as, in some cases, clinical tumours with associated histopathology measurements.

However, apart from physical accuracy, different characteristics can be equally important in terms of assessing the performance of such advanced image segmentation algorithms, which in principle have the potential to be more robust and repeatable than "threshold-based" approaches. A robust and repeatable performance may facilitate their use with images acquired on different scanner models without any previous optimization to individual image quality, providing a less hardware-dependent solution to the problem of 3-D functional uptake segmentation. However, none of these methodologies have been shown to be system-independent, considering the potential variability that can be observed in PET image characteristics depending on the scanner or associated reconstruction and correction algorithms used. Such an evaluation is essential for the efficient application of these approaches to the different clinical applications targeted, not simply within a given institution but also with regard to their use within a multicentre trial context.

Finally, such a robustness analysis could provide some insight into the potential behaviour of a given segmentation algorithm with the use of different tracers. On the one hand, the PET scanner properties in terms of spatial resolution will be similar for acquisitions performed with the same radioisotope, therefore resulting in partial volume effects of similar magnitude. On the other hand, acquisitions performed using different radiotracers show different uptake intensities and therefore subsequent different contrast and noise level characteristics for a given tumour uptake. For instance, $^{18}$F-FLT and $^{18}$F-FMISO images are usually characterized by higher noise levels and reduced tumour uptake contrast than $^{18}$F-FDG images [17, 18]. Therefore, studying the behaviour of automated algorithms dedicated to the delineation of elevated activity in $^{18}$F-FDG images, considering variable contrast and noise levels, could provide an insight into the potential behaviour of such algorithms when applied to other $^{18}$F-labelled PET tracers.

The objectives of this study were (1) to provide a robustness and repeatability evaluation framework, and (2) to assess within this framework the performance of different advanced and threshold-based segmentation algorithms in delineating elevated activity distributions in a PET image.

## Materials and methods

### Segmentation algorithms

Threshold-based and more advanced approaches were considered in this work. Two different fixed thresholds were considered, at 42% (T42) and 50% (T50) of the maximum tumour value, using a region growing algorithm with the maximum intensity voxel as seed [4]. An adaptive thresholding (TSBR, for threshold source-to-background ratio) approach [7] was also included:

$$I_{threshold} = a + b\frac{1}{SBR} \quad (1)$$

where SBR is the tumour-to-background ratio determined by ROI analysis, and the parameters $a$ and $b$ are optimized for each scanner using phantom acquisitions of spheres.

In terms of more advanced image segmentation approaches, the fuzzy C-means (FCM) [16] clustering, previously used for functional volume segmentation tasks in both brain and oncology applications [14, 15, 19, 20], was considered. This algorithm iteratively estimates cluster "centroids" (centres of mass) in the image, computing a voxel's membership between 0 and 1 to a given cluster depending on the distance between the voxel's value and the cluster centroids. However, the FCM algorithm lacks explicit noise and spatial correlation modelling. The second advanced algorithm considered was an unsupervised bayesian segmentation, known as the fuzzy locally adaptive bayesian (FLAB) algorithm [14, 15]. It computes, for each

voxel, a probability of belonging to a given "class" (for instance, tumour, background or a given uptake level within a tumour). This probability takes into account the voxel intensity, spatial correlation with surrounding voxels (the assumption being that voxels of similar intensities and close to each other have higher probability of belonging to the same class) as well as the overall statistical distributions in the regions of the image by estimating the mean and variance for each class. The FLAB algorithm automatically estimates the parameters of interest (number of classes, class mean and variance, spatial correlation of each voxel) within a stochastic expectation maximization (SEM) framework [21].

In order to deal with the inherent blurry properties of PET images due to the limited spatial resolution of scanners, the algorithm considers that each voxel may contain a mixture of classes by modelling both spatial correlation and statistical distributions with a combination of Dirac "hard" and Lebesgue "fuzzy" measures. This enables a classification of voxels as belonging to what we denote as "hard classes" or "fuzzy transitions", the first referring to fairly homogeneous regions, the second to blurred areas occurring at the frontier between two homogeneous regions. The FLAB algorithm is therefore able to accurately differentiate if necessary both the overall tumour spatial extent from its surrounding background as well as tumour subvolumes with different uptakes. The accuracy of the FLAB algorithm has been previously extensively investigated for both homogeneous [14] and heterogeneous nonspherical tumours [15] and has demonstrated satisfactory accuracy even for small (<2 cm diameter) volumes of interest (both overall tumours and tumour subvolumes), short acquisition durations (associated with higher noise levels) and low (<4:1) contrast (both for overall tumours with respect to their surrounding background and between a tumour and its smaller subvolumes).

## Accuracy, robustness, repeatability: definitions

For a given segmentation algorithm we define accuracy as the precision in retrieving the true 3-D object spatial extent, shape and volume based on the reconstructed activity distribution in a PET image, irrespective of the correlation between this distribution and the underlying physiological process. Thus an image segmentation algorithm would not be expected to differentiate specific from nonspecific tracer uptake (for example inflammation and tumour in the case of FDG) if they are of the same intensity. The defined accuracy of each of the methodologies considered was determined as in previous studies [14, 15] by calculating the classification errors (see section Analysis).

We define robustness as the ability of a given methodology to generate accurate segmented volumes under varying acquisition and image reconstruction conditions. This robustness is determined as the variability of the segmentation results when a method is applied without prior optimization to images acquired using various scanners, and for each scanner under various contrast and noise conditions, using different reconstruction and associated correction algorithms. A dataset consisting of multiple phantom acquisitions performed on various scanner models (see section Validation studies) was used for this task. These phantom studies were used to assess robustness as they are consistently employed for optimization purposes with most of the functional volume segmentation algorithms.

Within the context of this study, repeatability is defined as the ability of a given algorithm to reach the same result when applied multiple times to a single image. In such a task, deterministic fixed threshold approaches will always give the same result. On the other hand, more advanced methods may give different results when applied multiple times to the same image. For example, adaptive thresholding segmentation may depend on a manually drawn background ROI and may thus result in variable delineations depending on the choice of this ROI. Finally, manual delineation may be considered as the least repeatable, even when considering a single operator (intraoperator variability). In order to compare the performances of the different segmentation algorithms in terms of repeatability, we used a series of simulated tumour images [22], as well as 15 different clinical cases (see section Validation studies).

## Validation studies

Four different PET/CT scanners currently used in clinical practice were used for the robustness study: namely, the Philips Gemini and Gemini TF (Philips Medical Systems, Cleveland, OH), the Siemens Biograph (SIEMENS Medical Solutions, Knoxville, TN) and the GE Discovery LS (GE Healthcare, Milwaukee, WI). In each case, scans of the IEC phantom containing spheres of various diameters (10, 13, 17, 22, 28, 37 mm) filled with $^{18}$F and placed on a hot uniform background were acquired. A standard protocol was designed to generate the following acquisitions for each scanner model: (a) two different SBRs (4:1 and 8:1), (b) three different scan durations (1, 2 and 5 min) to study the effect of noise, and (c) two different voxel volumes used in the reconstruction (between 2×2×2 mm and 4.3×4.3×4.25 mm). All scans were performed in 3-D mode and list-mode format facilitating the generation of 1-, 2- and 5-min realizations from one single 5-min acquisition. In addition to the standard CT acquisition used for attenuation correction, a CT scan at high resolution was acquired for each PET/CT acquisition in order to generate (after registration) a ground-truth defining the true spatial extent

**Table 1** Overview of the acquisition parameters used for each scanner model

| PET/CT system | Contrast | Voxel size (mm) | Duration (min) | Reconstruction protocol |
|---|---|---|---|---|
| Philips Gemini | 4:1<br>8:1 | 2×2×2<br>4×4×4 | 1, 2, 5 | RAMLA 3D |
| Philips Gemini TF | 4:1<br>8:1 | 2×2×2<br>4×4×4 | 1, 2, 5 | TF ML-EM |
| Siemens Biograph | 4:1<br>8:1 | 2×2×2<br>5.33×5.33×2 | 1, 2, 5 | FORE-OSEM |
| GE Discovery LS | 4:1<br>8:1 | 1.95×1.95×4.25<br>4.3×4.3×4.25 | 1, 2, 5 | FORE-OSEM |

(the interior of the sphere) of the tracer uptake at the voxel-by-voxel level [14]. This is subsequently used to compute the accuracy of each algorithm through classification errors (see section Analysis).

Routine clinical image reconstruction protocols were used for all scanners. For the Philips GEMINI and GEMINI TF, data were reconstructed using the RAMLA 3D (two iterations, relaxation parameter 0.05, and 5-mm FWHM 3-D gaussian postfiltering) and the TF ML-EM algorithm, respectively. In the case of the Siemens Biograph and GE Discovery LS, images were reconstructed with Fourier rebinning (FORE) followed by OSEM (four iterations and eight subsets, with 5-mm FWHM 3-D gaussian postfiltering, and two iterations and eight subsets, respectively). All acquisitions were corrected for attenuation (using the corresponding CT image), as well as for scatter and random coincidences. A summary of the parameters for each of the datasets obtained using the different scanners is shown in Table 1. Figures 1 and 2 illustrate the various images obtained. Note that in the case of the Philips GEMINI acquisitions, the 37-mm sphere was not in the same plane as the others, and thus appears visually smaller in the selected slice, while the 28-mm sphere was missing in the phantom used for the GE Discovery LS acquisitions.

Regarding the repeatability study, two different datasets were used. The first one consisted of ten tumours extracted from a database of realistically simulated PET scans based on clinical whole-body images using the NCAT (NURBS cardiac-torso) phantom, a model of the Philips GEMINI

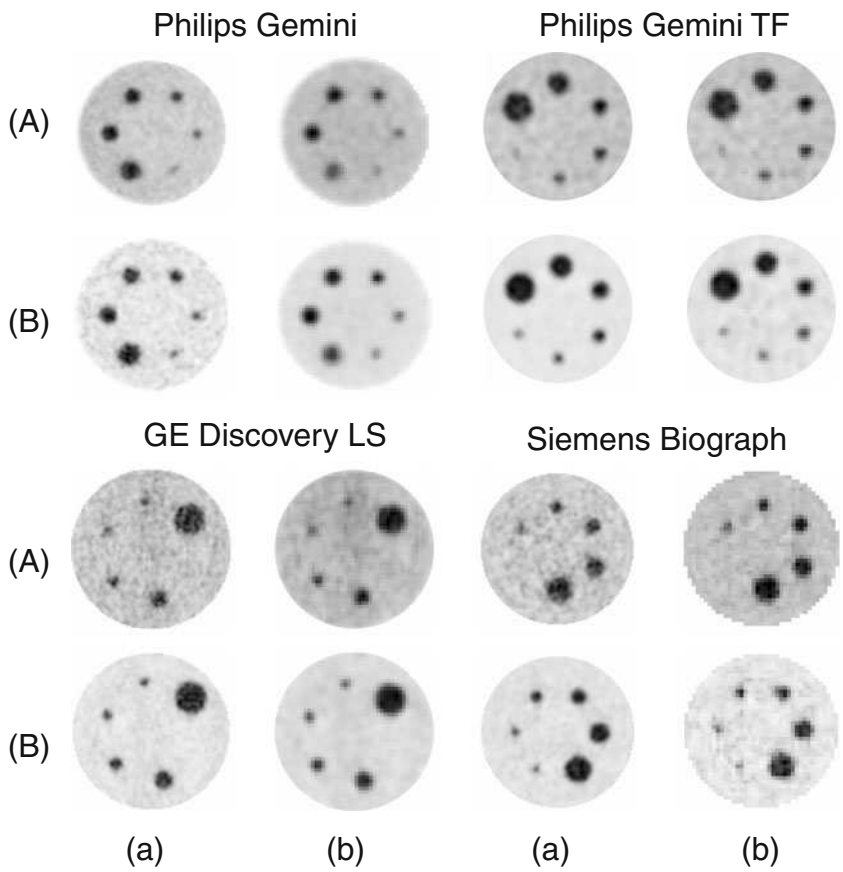



**Fig. 1** 2-D phantom slices through the centre of the spheres for the different systems and imaging conditions. Contrast ratios: rows *(A)* 4:1, rows *(B)* 8:1. Voxel sizes: columns *(a)* small voxels, columns *(b)* large voxels (see Table 1)

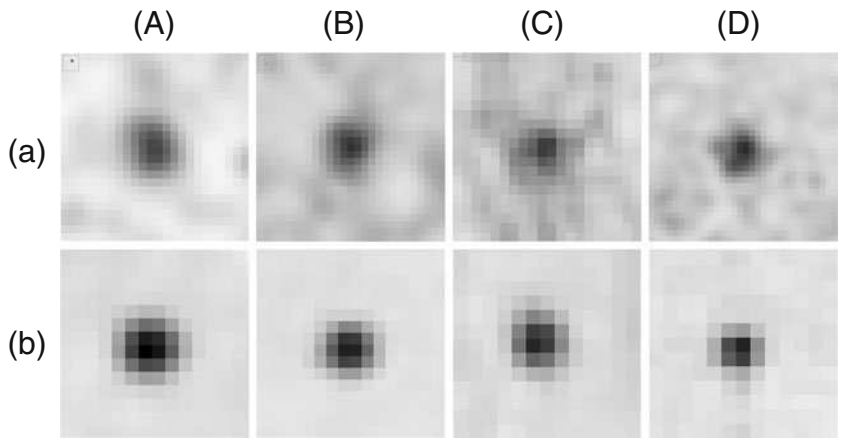



**Fig. 2** Variability for the 17-mm sphere across all four scanner models for two different configurations. Rows *(a)*: contrast 4:1, small voxels and 1-min acquisition. Rows *(b)*: contrast 8:1, large voxels and 5 min acquisition. Columns *(A)* Philips Gemini, *(B)* Philips Gemini TF, *(C)* Siemens Biograph, and *(D)* GE Discovery LS

scanner and GATE (Geant4 Application for Tomography Emission). The procedure for the generation of these images, reconstructed using OPL-EM (seven iterations, one subset) with 4×4×4 $mm^3$ voxels, has been previously described in detail [22]. In the second part of the repeatability study a number of clinical cases were selected from datasets acquired on various scanner models: four oesophagus lymphomas and four follicular lymphomas were scanned on the Philips GEMINI PET/CT scanner (2 min per bed position, 60 min after injection of 6 MBq/kg $^{18}$F-FDG); and three non-small-cell lung cancers were scanned on the Siemens Biograph (5 min per bed position, 45 min after injection of 5 MBq/kg $^{18}$F-FDG) and on the GE Discovery LS (3 min per bed position, 60 min after injection of 5 MBq/kg $^{18}$F-FDG).

## Analysis

For the phantom images used in the robustness study each sphere was processed separately. The images corresponding to the region containing each sphere were segmented in two classes (*sphere* and *background*), using each of the methods under evaluation (FCM, FLAB, T42 and T50). A voxel-to-voxel ground-truth based on the corresponding CT datasets as described previously [14] was used in the robustness evaluation of the different methodologies considered through the determination of the segmentation accuracy with the computation of the classification errors (CE):

$$CE = \frac{card\{t|c_t \neq x_t\}}{card\{t|x_t = 1\}} \times 100 \quad (2)$$

where, $c_t$ is the class assigned by the classification of voxel $t$, and $x_t$ is its true class ($x_t$=1 for the sphere and $x_t$= 0 for the background) and card{} is the cardinal. The errors are computed based on all misclassified voxels, either background voxels classified as sphere voxels or vice versa, divided by the total number of voxels defining the sphere volume.

The mean classification error and associated standard deviation (SD) were obtained for each sphere and for each segmentation approach, thus providing a measure of the robustness of the different segmentation algorithms when applied without specific optimization for a given scanner model or associated reconstruction algorithm under different imaging conditions (contrast and noise). The 10-mm sphere was not included in the analysis because it was not clearly visible in several of the phantom acquisitions and was therefore not possible to segment particularly when using 4×4×4 $mm^3$ and 5×5×5 $mm^3$ reconstruction voxel sizes by any of the segmentation algorithms considered. Adaptive thresholding could not be compared directly with the other methodologies since it is optimized on each of the

**Table 2** Optimized parameters *a* and *b* of the adaptive thresholding (TSBR) approach for each scanner model, with the minimum mean classification errors and their associated standard deviations across the entire range of configurations

| PET/CT system | TSBR approach | | Minimum mean associated classification error (%) | Standard deviation of classification error |
|---|---|---|---|---|
| | Parameter *a* | Parameter *b* | | |
| Philips Gemini | 40.1 | 59.7 | 10.8 | 3.3 |
| Philips Gemini TF | 38.6 | 61.4 | 9.7 | 2.8 |
| Siemens Biograph | 41.7 | 57.6 | 13.1 | 5.2 |
| GE Discovery LS | 42.0 | 56.8 | 11.1 | 3.7 |

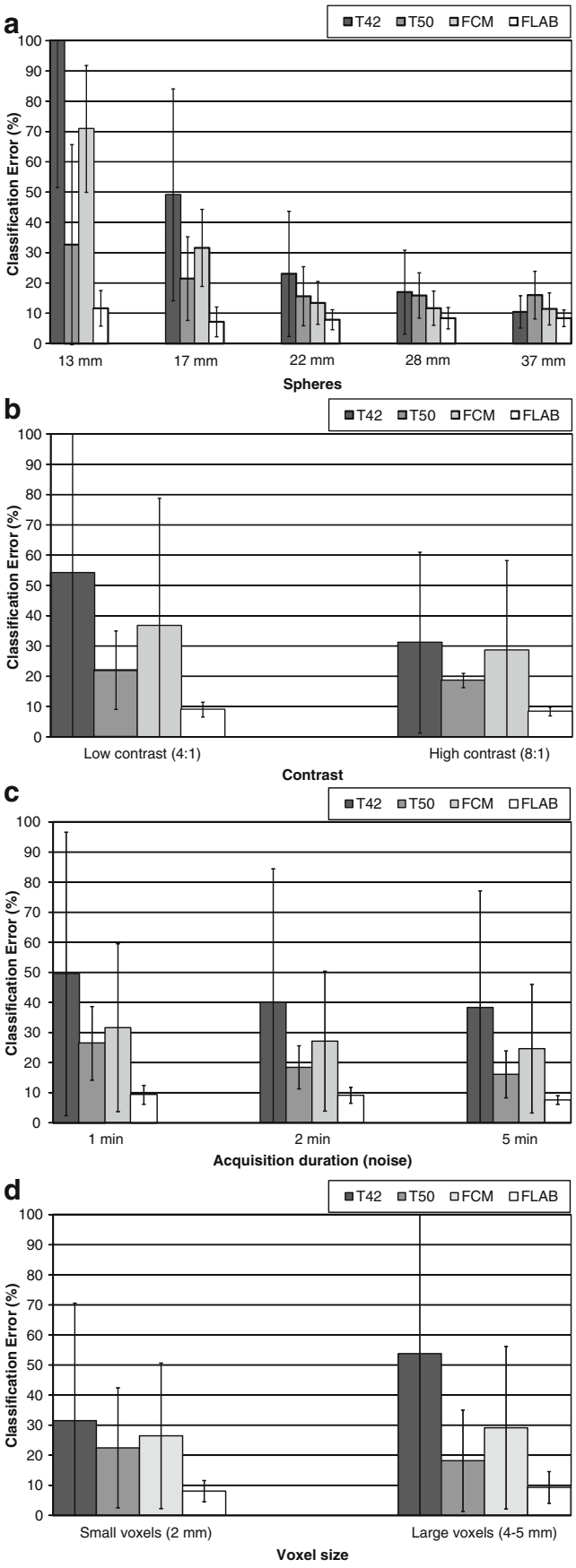



**Fig. 3** Mean classification errors and standard deviations (error bars) for each methodology with respect to (**a**) sphere diameter, (**b**) contrast, (**c**) acquisition duration and (**d**) voxel size, computed across the different scanner models

individual scanner datasets, with the parameters *a* and *b* optimized for each imaging device shown in Table 2. However, in order to assess the robustness of these approaches depending on the imaging system used we applied adaptive thresholding using the parameters optimized on other scanners to the image datasets acquired with the Siemens Biograph.

For the repeatability evaluation, the simulated and clinical tumours were segmented ten times each with the FCM, FLAB and TSBR algorithms (fixed thresholding was not included since it always gives the same volume). In addition, manual delineation was carried out by two nuclear medicine experts with similar experience (more than 10 years) and training. More specifically the two experts were instructed to delineate the elevated uptakes in the images by performing ten different slice-by-slice manual delineations for the different lesions considered in a randomized fashion, ensuring a minimum of 1 week between two consecutive segmentations of the same lesion. All these manual segmentations were carried out under the same conditions of full range contrast display. The mean percentage variability and associated standard deviation with respect to the mean segmented volume was computed for each of the lesions and segmentation approaches across the ten executions and across the ten manual delineations in order to assess the repeatability of the approaches for each of the images. The repeatability of the manual delineations from the two experts were compared separately (intra-observer variability) and with each other (interobserver variability).

## Results

Classification errors representing segmentation accuracy computed for each sphere are shown in Fig. 3a, considering the entire range of systems used for acquisition and the different parameters in terms of contrast, acquisition duration and voxel size. For all the systems considered, the relative impact of the different acquisition (contrast, duration) and reconstruction (voxel size) parameters is demonstrated in Fig. 3b, c and d, respectively. Table 3 shows the mean errors and standard deviations computed across the different spheres taken separately (as shown in Fig. 3a) and all together for the different imaging devices and acquisition configurations considered.

For the entire range of sphere sizes (37 to 13 mm), the FLAB algorithm showed better accuracy and variability through smaller overall mean errors and SD (8.7±4.5%)

**Table 3** Robustness evaluation: mean classification error and associated standard deviation computed for each methodology across the entire range of sphere phantom acquisitions

| Sphere diameter (mm) | Classification error (%) | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | T42 | | T50 | | FCM | | FLAB | |
| | Mean | SD | Mean | SD | Mean | SD | Mean | SD |
| 37–13 (all spheres) | 42.6 | 51.6 | 20.3 | 18.5 | 27.8 | 25.6 | 8.7 | 4.5 |
| 37 | 10.5 | 5.3 | 16 | 7.9 | 11.4 | 5.3 | 8.4 | 2.8 |
| 28 | 17 | 13.8 | 15.9 | 7.5 | 11.7 | 5.7 | 8.4 | 3.6 |
| 22 | 23 | 20.7 | 15.6 | 9.8 | 13.4 | 7.1 | 7.9 | 3.3 |
| 17 | 49.1 | 35 | 21.5 | 13.8 | 31.6 | 12.7 | 7.2 | 4.9 |
| 13 | 113.6 | 62.1 | 32.7 | 33.1 | 70.9 | 20.9 | 11.6 | 5.9 |

than the other advanced segmentation algorithm FCM (27.8±25.6%) as well as relative to the fixed threshold approaches T50 (20.3±18.5%) and T42 (42.6±51.6%). These latter were also more sensitive to variations in the parameters as shown in Fig. 3a. The T50 algorithm was clearly more robust than T42 algorithm (SD 19% compared to 52%). This is explained by the fact that the 50% threshold is more restrictive and hence leads to lower over-estimation for the smallest sphere volumes, and that the 42% threshold may lead to a gross over-estimate (>100% errors for the most challenging imaging conditions). On the other hand, the T50 algorithm was associated with a larger classification error for the two larger spheres, as it tended to under-estimate their volumes by only including the central high-intensity voxels of the sphere. The FCM algorithm was unable to accurately segment spheres smaller than 2 cm in diameter, leading to large overall mean errors when considering its performance over all sphere sizes, although it exhibited a lower variability than the fixed threshold approaches for the majority of the spheres with a size of >2 cm.

As shown in Fig. 3b, the FLAB algorithm exhibited low variability with respect to contrast changes, and all other methodologies, especially the T42 and FCM algorithms, exhibited higher sensitivity to such changes. The T50 algorithm, on the other hand, was less sensitive to contrast changes with respect to the mean error but exhibited larger variability for lower contrast. Figure 3c illustrates the resilience to shorter acquisitions (hence higher noise levels) for each methodology. The FLAB algorithm demonstrates very low variability with shorter acquisitions, whereas all other methodologies showed higher variability with significantly larger mean errors and standard deviations. Finally, only small improvements were seen for each methodology (except for T50) when using smaller voxels (see Fig. 3d).

The optimized parameters $a$ and $b$ of the TSBR approach for each scanner model are shown in Table 2. The mean classification error across all the spheres (range 13–37 mm) associated with each scanner was between 9.7% and 13.1% with associated standard deviations from 2.8% to 5.2%. When applying the parameters $a$ and $b$ of the Philips GEMINI, Philips GEMINI TF and Discovery LS datasets to the Siemens Biograph dataset, the mean error increased from 13.1±5.2% to 21.7±7.1%, 23.4±7.6% and 19.1±6.4%, respectively.

Concerning repeatability, Table 4 shows the mean variability and SD around the mean segmented volume across the ten manual delineations performed by the two nuclear medicine experts, and ten repeated executions of

**Table 4** Repeatability evaluation: variability and standard deviation around the mean segmented volume for repeated (10 times) delineations of simulated and clinical tumours

| Method | Variability (%) | | | |
|---|---|---|---|---|
| | Simulated cases | | Clinical cases | |
| | Mean | SD | Mean | SD |
| FLAB | 0.5 | 0.3 | 0.9 | 0.5 |
| FCM | 0.8 | 0.6 | 1.7 | 1.9 |
| Fixed thresholding | 0 | 0 | 0 | 0 |
| Adaptive thresholding | 3.4 | 2.8 | 3.8 | 3.1 |
| Manual delineation | | | | |
| Expert 1 | 13.4 | 17.3 | 19.6 | 15.2 |
| Expert 2 | 11.7 | 18.4 | 22.1 | 13.6 |
| Expert 2 with respect to expert 1 | 16.4 | 21.8 | 24.7 | 17.5 |

the FLAB, FCM and TSBR algorithms. The FLAB algorithm demonstrated highly repeatable results in all of the studied cases, with negligible variability (<1%) around the mean segmented 3-D volume across the different repeated executions for both the simulated and the clinical datasets. The FCM algorithm also led to satisfactory repeatability results (0.8±0.6% for the simulated tumours and 1.7±1.9% for the clinical cases). However, the variability with the TSBR algorithm was more than double (3.4±2.8% for the simulated tumours and 3.8±3.1% for the clinical cases) which was most probably due to the manual definition of the background ROI. By contrast manual segmentation performed by the two experts showed high intraobserver variability for simulated tumours (13.4±17.3% and 11.7±18.4% for expert 1 and 2, respectively), and even larger variability for the clinical images (19.6±15.2% and 22.1±13.6% for expert 1 and 2, respectively). Interobserver variability was 16.4±21.8% and 24.7±17.6% for the simulated tumours and clinical cases, respectively. Figure 4 shows examples of delineations obtained by manual segmentation and the automatic approaches.

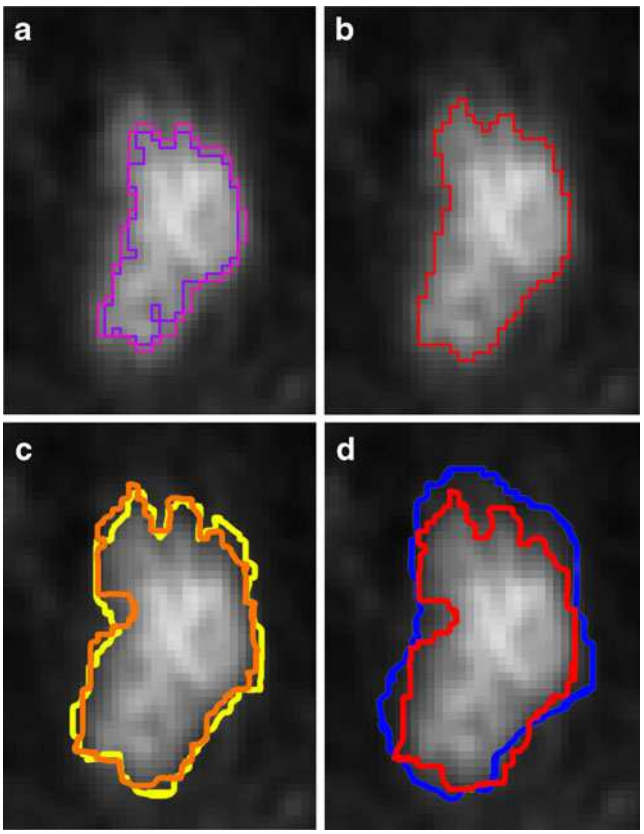



**Fig. 4** Tumour delineations on the same image slice: **a** delineation by adaptive thresholding with two different background ROIs (6% difference); **b** delineation by the FLAB method; **c** two manual delineations by the same observer (fairly consistent, 9% difference); **d** two manual delineations by different observers (highly different, 37% difference)

## Discussion

Functional tumour uptake volume delineation is today an area of interest for multiple clinical (routine and research) applications of PET imaging, such as studying response to therapy and radiotherapy treatment planning. In all of these applications, the robustness and repeatability with which functional uptake volumes can be determined under different imaging conditions play a predominant role in allowing the level of confidence to be established with the use of such tumour volume measurements in clinical practice [18]. Although several promising advanced algorithms have recently been proposed [11–15, 20], methodologies currently used in clinical practice are based on the use of manual delineation or fixed and adaptive thresholding [6–8]. The major drawback of manual delineation is its high inter- and intraobserver variability, in addition to being time consuming. On the other hand, the currently considered state-of-the-art adaptive threshold-based algorithms have been shown to accurately define functional volumes under certain imaging conditions of spherical lesions with a homogeneous activity distribution. However, they require specific parameter optimization and are thus system-dependent. In addition, the adaptive thresholding approaches usually involve some user interaction to select background regions of interest, which can potentially lead to user-introduced variability. In the present study we focused on the evaluation under different imaging conditions of the level of robustness and repeatability of different functional volume segmentation algorithms, including those used in current clinical practice.

In terms of robustness, the use of images from different commercial PET/CT systems acquired under typical clinical acquisition conditions resulted in large variability in the performance of the different segmentation algorithms evaluated. Across all of the images and spheres considered, a fixed threshold of 42% of the maximum resulted in the largest variability of the segmented functional volumes (±15–60%) across the different images considered for spheres <3 cm in diameter. On the other hand, the variability using a fixed threshold of 50% was closer (±20%) to that of one of the advanced segmentation algorithms included in this work (FCM). Finally, the FLAB algorithm was the most robust of all the evaluated algorithms leading to the lowest variability (±5%), with no particular dependence on acquisition (duration, contrast) and processing parameters (reconstructed voxel size). The 42% fixed threshold and the FCM algorithm were the most sensitive to contrast and acquisition duration across the different scanners used. In terms of variability across the different images used, the 50% fixed threshold demonstrated the most significant dependence of variability on lesion contrast. Finally, applying adaptive thresholding (TSBR) to

acquisitions performed on a different scanner than the one used to optimize its parameters led to higher mean errors of <25%.

In terms of repeatability, all algorithms considered exhibited mean differences of <5%, although only the FLAB algorithm came close to the perfect repeatability that can be achieved by a deterministic approach such as a fixed threshold. Finally, the repeatability of both threshold and automatic segmentation approaches was superior to that of manual delineation (variability >15–20% for both the clinical and simulated tumours).

The overall better accuracy (lower mean errors) and lower variability (lower standard deviations) associated with the FLAB algorithm across the different images considered demonstrates its ability, without the need of any scanner-specific optimization, to robustly deal with the different image qualities resulting from the use of different reconstruction and correction algorithms as well as sensitivities associated with different systems. This of course should be considered within the context of the limited absolute accuracy of binary threshold-based approaches shown in this and previous studies. The accuracy of threshold-based approaches is particularly limited for lesions with a nonhomogeneous form and activity distribution. In such cases it may result, as previously shown [15], in large under- or over-estimation of the overall tumour spatial extent.

The present study also demonstrated that the use of any of the segmentation algorithms significantly reduced intra- and interobserver variability associated with manual delineation. However, one should keep in mind that automated segmentation algorithms are not able to differentiate between similar levels of physiological and pathological elevated tracer uptakes. Therefore physician involvement is still imperative and desirable, especially regarding the detection and selection of elevated tracer uptakes corresponding to pathological findings that are to be subsequently accurately delineated.

## Conclusion

This study demonstrated significant differences in the robustness and reproducibility of functional volume measurements depending on the segmentation algorithm used. The advantage of employing advanced segmentation algorithms is an improvement in overall elevated activity delineation across the range of image qualities that can be encountered today in clinical practice, without the need for system-dependent optimization procedures. In addition, their high level of repeatability allows performance similar to that of deterministic threshold-based approaches to be achieved. Therefore such advanced image segmentation algorithms may provide robust and reliable tools to aid physicians as an initial guess in determining functional volumes on PET images.

**Acknowledgments** This work was financially supported by the French National Research Agency (ANR) under contract ANR-08-ETEC-005-01. We would like to thank the following clinical centres and associated members for some of the phantom and patient datasets used in this study: the nuclear medicine departments of CHU Brest, France (Alexandre Turzo), CHU Sud-Amiens, France (Pascal Bailly, Joel Daouk), and St Bartholomew's Hospital, London, UK (Iain Murray).

**Conflicts of interest** None

## References

1. Lucignani G. SUV and segmentation: pressing challenges in tumor assessment and treatment. Eur J Nucl Med Mol Imaging. 2009;36:715–20.
2. Jarritt H, Carson K, Hounsel AR, Visvikis D. The role of PET/CT scanning in radiotherapy planning. Br J Radiol. 2006;79(S):27–35.
3. Pan T, Mawlawi O. PET/CT in radiation oncology. Med Phys. 2008;35(11):4955–66.
4. Krak NC, Boellaard R, Hoekstra OS, Twisk JW, Hoekstra CJ, Lammertsma AA, et al. Effects of ROI definition and reconstruction method on quantitative outcome and applicability in a response monitoring trial. Eur J Nucl Med Mol Imaging. 2005;32:294–301.
5. Jerusalem G, Hustinx R, Beguin Y, Fillet G. The value of positron emission tomography (PET) imaging in disease staging and therapy assessment. Ann Oncol. 2002;13(S4):227–34.
6. Erdi YE, Mawlawi O, Larson SM, Imbriaco M, Yeung H, Finn R, et al. Segmentation of lung lesion volume by adaptive positron emission tomography image thresholding. Cancer. 1997;80(12 Suppl):2505–9.
7. Daisne JF, Sibomana M, Bol A, Doumont T, Lonneux M, Grégoire V. Tri-dimensional automatic segmentation of PET volumes based on measured source-to-background ratios: influence of reconstruction algorithms. Radiother Oncol. 2003;69:247–50.
8. Nestle U, Kremp S, Schaefer-Schuler A, Sebastian-Welsch C, Hellwig D, Rübe C, et al. Comparison of different methods for delineation of 18F-FDG PET-positive tissue for target volume definition in radiotherapy of patients with non-Small cell lung cancer. J Nucl Med. 2005;46(8):1342–8.
9. Biehl KJ, Kong MF, Dehdashti F, Jin JY, Mutic S, El Naqa I, et al. 18F-FDG PET definition of gross tumor volume for radiotherapy of non-small cell lung cancer: is a single standardized uptake value threshold approach appropriate? J Nucl Med. 2006;47:1808–12.
10. Oellers M, Bosmans G, van Baardwijk A, Dekker A, Lambin P, Teule J, et al. The integration of PET-CT scans from different hospitals into radiotherapy treatment planning. Radiother Oncol. 2008;87(1):142–6.
11. El Naqa I, Yang D, Apte A, Khullar D, Mutic S, Zheng J, et al. Concurrent multimodality image segmentation by active contours for radiotherapy treatment planning. Med Phys. 2007;34 (12):4738–49.
12. Montgomery DW, Amira A, Zaidi H. Fully automated segmentation of oncological PET volumes using a combined multiscale and statistical model. Med Phys. 2007;34(2):722–36.
13. Geets X, Lee JA, Bol A, Lonneux M, Grégoire V. A gradient-based method for segmenting FDG-PET images: methodology and validation. Eur J Nucl Med Mol Imaging. 2007;34:1427–38.

14. Hatt M, Cheze le Rest C, Turzo A, Roux C, Visvikis D. A fuzzy Bayesian locally adaptive segmentation approach for volume determination in PET. IEEE Trans Med Imaging. 2009;28 (6):881–93.
15. Hatt M, Cheze le Rest C, Descourt P, Dekker A, De Ruysscher D, Oellers M, et al. Accurate automatic delineation of heterogeneous functional volumes in positron emission tomography for oncology applications. Int J Radiat Oncol Biol Phys. 2010;77(1):301–8.
16. Dunn JC. A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. J Cybernet. 1974;31:32–57.
17. Koh WJ, Rasey JS, Evans ML, Grierson JR, Lewellen TK, Graham MM, et al. Imaging of hypoxia in human tumors with [F-18]fluoromisonidazole. Int J Radiat Oncol Biol Phys. 1992;22 (1):199–212.
18. Hatt M, Cheze Le Rest C, Aboagye EO, Kenny LM, Rosso L, Turkheimer FE, et al. Reproducibility of 18F-FDG and 3'-deoxy-3'-18F-fluorothymidine PET tumor volume measurements. J Nucl Med. 2010;51(9):1368–76.
19. Zhu W, Jiang T. Automation segmentation of PET image for brain tumors. IEEE Nucl Sci Symp Conf Rec. 2003;4:2627–9.
20. Belhassen S, Zaidi H. A novel fuzzy C-means algorithm for unsupervised heterogeneous tumor quantification in PET. Med Phys. 2010;37(3):1309–24.
21. Celeux G, Diebolt J. L'algorithme SEM: un algorithme d'apprentissage probabiliste pour la reconnaissance de mélange de densités. Rev Statist Appl. 1986;34(2):35–52.
22. Le Maitre A, Segars WP, Marache S, Reilhac A, Hatt M, Tomei S, et al. Incorporating patient specific variability in the simulation of realistic whole body 18F-FDG distributions for oncology applications. Proc IEEE. 2009;97(12):2026–38.

Eur J Nucl Med Mol Imaging
DOI 10.1007/s00259-011-1755-7

ORIGINAL ARTICLE

# Prognostic value of $^{18}$F-FDG PET image-based parameters in oesophageal cancer and impact of tumour delineation methodology

**Mathieu Hatt · Dimitris Visvikis · Nidal M. Albarghach · Florent Tixier · Olivier Pradier · Catherine Cheze-le Rest**

Received: 6 October 2010 / Accepted: 1 February 2011


**Abstract**
*Purpose* $^{18}$F-fluorodeoxyglucose (FDG) positron emission tomography (PET) image-derived parameters, such as standardized uptake value (SUV), functional tumour length (TL) and tumour volume (TV) or total lesion glycolysis (TLG), may be useful for determining prognosis in patients with oesophageal carcinoma. The objectives of this work were to investigate the prognostic value of these indices in oesophageal cancer patients undergoing combined chemo-radiotherapy treatment and the impact of TV delineation strategies.
*Methods* A total of 45 patients were retrospectively analysed. Tumours were delineated on pretreatment $^{18}$F-FDG scans using adaptive threshold and automatic (fuzzy locally adaptive Bayesian, FLAB) methodologies. The maximum standardized uptake value ($SUV_{max}$), $SUV_{peak}$, $SUV_{mean}$, TL, TV and TLG were computed. The prognostic value of each parameter for overall survival was investigated using Kaplan-Meier and Cox regression models for univariate and multivariate analyses, respectively.
*Results* Large differences were observed between methodologies (from −140 to +50% for TV). SUV measurements were not significant prognostic factors for overall survival, whereas TV, TL and TLG were, irrespective of the segmentation strategy. After multivariate analysis including standard tumour staging, only TV ($p<0.002$) and TL ($p=0.042$) determined using FLAB were independent prognostic factors.
*Conclusion* Whereas no SUV measurement was a significant prognostic factor, TV, TL and TLG were significant prognostic factors for overall survival, irrespective of the delineation methodology. Only functional TV and TL derived using FLAB were independent prognostic factors, highlighting the need for accurate and robust PET tumour delineation tools for oncology applications.

**Keywords** PET · Tumour volume · Tumour segmentation · Oesophageal cancer · Survival

M. Hatt (✉) · D. Visvikis · N. M. Albarghach · F. Tixier · O. Pradier · C. Cheze-le Rest
INSERM, U650 LaTIM, CHU Morvan,
5 Avenue Foch,
29609 Brest, France
e-mail: hatt@univ-brest.fr

C. Cheze-le Rest
Academic Department of Nuclear Medicine, CHU Morvan,
5 Avenue Foch,
29609 Brest, France

N. M. Albarghach · O. Pradier
Department of Radiotherapy, CHU Morvan,
5 Avenue Foch,
29609 Brest, France

## Introduction

The incidence of oesophageal cancer is increasing and despite advances in therapy, the diagnosis still carries a poor prognosis [1]. As with all tumours, the outcome for patients is strongly associated with the stage at initial diagnosis. The TNM (tumour, node, metastasis) system currently in use for the staging of oesophageal cancer does not take into account non-anatomical factors such as histopathological type, grade or various biomarkers that may be important determinants of prognosis. The pathological stage is given by surgery but this procedure is not possible for all patients because it is associated with a high risk of mortality and morbidity. Therefore a noninvasive

Published online: 02 March 2011

staging method would be of great interest, and within this context the primary role of $^{18}$F-fluorodeoxyglucose (FDG) positron emission tomography (PET) in oesophageal cancer is the detection of distant metastases [2–4]. This modality is also gaining acceptance in oesophageal cancer for the assessment of therapy response [5, 6] or for radiotherapy treatment planning [7–9]. Lately, some authors have also suggested that different parameters derived from initial $^{18}$F-FDG PET images could have a role as independent prognostic factors [10–14]. The parameters studied include standardized uptake value ($SUV_{max}$ as the maximum uptake in the primary tumour or in the combined primary and regional area), tumour functional longitudinal length (TL), tumour functional volume (TV), nodal uptake or FDG-avid metastases [10–14]. Although a few studies have demonstrated the interest of these indices for determining prognosis, there are conflicting results concerning the independent prognostic value of each of these indices. For example, Hyun et al. [12], analysing results from 151 patients with principally squamous cell carcinoma (SCC), have recently suggested that primary tumour $SUV_{max}$ is not an independent prognostic factor, in agreement with other studies [10, 15, 16]. On the other hand, Kato et al. [17] based on the analysis of 184 patients with oesophageal SCC have shown that $SUV_{max}$ of the primary tumour is an independent prognostic factor for overall survival, in agreement with other studies [18–20]. These conflicting results can be potentially caused by differences in the methodology used for the analysis of the PET images. Although $SUV_{max}$ is less sensitive to tumour size, the conflicting results considering its value as an independent prognostic factor may also be due to variability in the tumour sizes considered in the different studies.

Pathological TL has been shown to be an independent prognostic factor in oesophageal carcinoma [21]. Therefore, determining the functional TL in $^{18}$F-FDG PET images may be a good surrogate. The use of different thresholds for the determination of the functional TL in the existing studies may be responsible for the conflicting results regarding its value as a predictor of response to chemoradiotherapy [11, 22], while it has been shown to be an independent predictor in patients undergoing surgery [10]. On the other hand, functional TV may be more representative of overall tumour burden. The value of the functional TV has been only recently explored in a single study of patients with oesophageal carcinoma considering a heterogeneous treatment regime (76 and 24% treated by surgery and combined radiochemotherapy, respectively) [12]. In this study both the presence of metastases and the TV were found to be independent prognostic factors for overall patient survival. Tumours were delineated based on a three fixed threshold scale depending on the tumour $SUV_{max}$. Although such an approach may be simple to implement in clinical practice, the use of a fixed threshold for functional TV determination suffers from multiple shortcomings which have been highlighted in different studies [23, 24]. In addition, the proposed scale is not universally applicable to the different clinical settings spanning from the acquisition protocols to the scanning systems used and variable associated image qualities.

Therefore, despite early evidence that functional TL and TV may be useful parameters in predicting survival and response to therapy, there is a clear need to assess the influence of the methodology used in obtaining these indices. Finally, the determination of functional TV may allow the calculation of alternative image-derived indices such as the total glycolytic lesion index (TLG) (g), defined as the product of the TV (ml) and its associated mean activity ($SUV_{mean}$) (g/ml) [25], whose value has not as yet been explored in predicting response to therapy or as a prognostic factor for survival using $^{18}$F-FDG in oesophageal carcinoma.

The objective of this study was therefore to retrospectively investigate the prognostic value of $^{18}$F-FDG PET in 45 oesophageal cancer patients treated with concomitant radiochemotherapy, considering for the first time in a single study all of the commonly used PET-derived parameters such as functional TL, TV, SUV measurements (max, peak, mean) and TLG. In addition, the impact of different tumour delineation strategies was assessed.

## Materials and methods

### Patients

A total of 45 patients with newly diagnosed oesophageal cancer treated between 2004 and 2008 with concomitant radiochemotherapy and without surgery were included in this study. The characteristics of the patients are given in Table 1. Of the patients, 41 were male (91%), and the mean age at the time of diagnosis was 66±10 years. Most of the tumours were SCC (73%) and originated from the middle and lower oesophagus (76%). All patients were referred before treatment for an $^{18}$F-FDG PET study as part of a routine procedure for the initial staging in oesophageal cancer. The treatment included three courses of 5-fluorouracil/cisplatin and a median radiation dose of 60 Gy given in 180-cGy daily fractions delivered once daily, 5 days a week for 6–7 weeks. Follow-up data were prospectively collected in a database for further analysis and overall survival was calculated. The current analysis was carried out after an approval by the Institutional Ethics Review Board.

### $^{18}$F-FDG PET acquisitions

$^{18}$F-FDG PET studies were carried out prior to the treatment. Patients were instructed to fast for a minimum

**Table 1** Patient demographic and clinical characteristics

| Parameter | No. of patients (%) |
|---|---|
| Gender | |
| Male | 41 (91) |
| Female | 4 (9) |
| Age | |
| Range | 45–84 |
| Median | 68 |
| Site | |
| Upper oesophagus | 11 (24) |
| Middle oesophagus | 17 (38) |
| Lower oesophagus | 17 (38) |
| Histology type | |
| Adenocarcinoma | 12 (27) |
| SCC | 33 (73) |
| Histology differentiation | |
| Well differentiated | 12 (27) |
| Moderately differentiated | 11 (24) |
| Poorly differentiated | 4 (9) |
| Unknown | 18 (40) |
| TNM stage | |
| T1 | 6 (13) |
| T2 | 7 (16) |
| T3 | 22 (49) |
| T4 | 10 (22) |
| N0 | 18 (40) |
| N1 | 27 (60) |
| M0 | 29 (64) |
| M1 | 16 (36) |
| AJCC stage | |
| I | 3 (7) |
| IIA | 7 (16) |
| IIB | 5 (11) |
| III | 14 (31) |
| IVa | 5 (11) |
| IVb | 11 (24) |

of 6 h before the injection of $^{18}$F-FDG. The administered dose was 5 MBq/kg, and static emission images were acquired (2 min per bed position) from head to thigh beginning 60 min after injection on a Philips GEMINI PET/CT system (Philips Medical Systems, Cleveland, OH, USA). In addition to the emission PET scan, a low-dose CT scan without IV or oral contrast was acquired for attenuation correction. Images were reconstructed with the 3-D row action maximum likelihood algorithm (RAMLA) using standard clinical protocol parameters: 2 iterations, relaxation parameter of 0.05, 5-mm 3-D Gaussian post-filtering and 4×4×4 mm$^3$ voxels grid sampling.

## PET image analysis

The following parameters were extracted from each PET image: $SUV_{max}$, $SUV_{peak}$ defined as the mean of $SUV_{max}$ and its 26 neighbours, mean SUV within the delineated tumour ($SUV_{mean}$), functional TL in longitudinal direction, functional TV and TLG. $SUV_{peak}$, usually defined as a 1-cm circle or sphere [26] [we used a 3×3×3 cube of 4×4×4 mm$^3$ voxels which roughly corresponds to the same region of interest (ROI)], was considered in order to investigate the impact of reducing the potential bias in the $SUV_{max}$ measurements as a result of its sensitivity to noise.

Whereas $SUV_{max}$ and $SUV_{peak}$ are independent on the tumour delineation strategy used, TL, TV, $SUV_{mean}$ and the derived TLG were determined on delineations performed using two strategies. First, an adaptive threshold [23] using a background ROI manually chosen by two experienced nuclear medicine physicians led to two different results $T_{bckgrd}^1$ and $T_{bckgrd}^2$. Observers were instructed to choose the ROI in the mediastinum at a sufficient distance from the lesion to avoid any overlapping. However, they were allowed to choose the size, shape and exact placement of the ROI. Finally the automatic fuzzy locally adaptive Bayesian (FLAB) algorithm [24, 27] was considered.

## Statistical analysis

All quantitative data were expressed as mean±1 standard deviation (SD) and summary statistics are given in Table 2.

The correlation between all parameters extracted using the different methodologies was computed with Pearson coefficients. The differences between methodologies were assessed using Bland-Altman analysis [28] to define bias as the mean error, and agreement intervals (upper and lower limits) as 1.96 times the SD. Kaplan-Meier methods were used to estimate the survival distributions [29]. Overall survival was calculated from the date of initial diagnosis to the date of death or most recent follow-up in cases of patients still alive. Survival curves were generated for each parameter considered. The most discriminating threshold value allowing differentiation of the two groups of patients was selected using receiver-operating characteristic (ROC) methodology [30]. The prognostic value of each parameter in terms of overall survival was assessed by the log-rank test. The significance of the following factors was tested: age, gender, histology type, T, N, M classifications, American Joint Committee on Cancer (AJCC) stage [31], TL, TV, $SUV_{max}$, $SUV_{peak}$, $SUV_{mean}$ and TLG. Independent prognostic factors for overall survival were determined using multivariate Cox regression analysis [32] by incorporating in the model all parameters that were deemed significant in the univariate analysis. However, the indices obtained by each delineation ($T_{bckgrd}^1$, $T_{bckgrd}^2$ and FLAB)

**Table 2** Parameters definition and statistics

| Definition | | | Notation | Mean ± SD | Range |
|---|---|---|---|---|---|
| Highest SUV within the tumour | | | $SUV_{max}$ | 10±3.8 | 2.2–19.7 |
| Mean of $SUV_{max}$ and its 26 neighbours | | | $SUV_{peak}$ | 8.2±3.3 | 1.8–16.1 |
| Mean SUV of tumour defined by | Adaptive threshold | 1st user | $SUV_{mean}$ ($T_{bckgrd}^1$) | 6.6±2.6 | 1.8–13.7 |
| | | 2nd user | $SUV_{mean}$ ($T_{bckgrd}^2$) | 6.2±2.7 | 1.6–13.8 |
| | FLAB | | $SUV_{mean}$ (FLAB) | 6.0±2.4 | 1.7–13.2 |
| Functional TV defined by | Adaptive threshold | 1st user | TV ($T_{bckgrd}^1$) | 22.6±23.8 | 1.8–92.0 |
| | | 2nd user | TV ($T_{bckgrd}^2$) | 29.2±29.7 | 2.4–133.9 |
| | FLAB | | TV (FLAB) | 36.3±33.7 | 3.0–139.7 |
| Functional TL defined by | Adaptive threshold | 1st user | TL ($T_{bckgrd}^1$) | 5.9±3.0 | 1.6–15.6 |
| | | 2nd user | TL ($T_{bckgrd}^2$) | 5.6±2.9 | 1.6–14.4 |
| | FLAB | | TL (FLAB) | 6.2±2.9 | 2.0–15.6 |
| $SUV_{mean}$ ($T_{bckgrd}^1$)×TV ($T_{bckgrd}^1$) (g) | | | TLG ($T_{bckgrd}^1$) | 165.4±182.7 | 3.2–759.7 |
| $SUV_{mean}$ ($T_{bckgrd}^2$)×TV ($T_{bckgrd}^2$) (g) | | | TLG ($T_{bckgrd}^2$) | 198.8±209.4 | 6.9–921.3 |
| $SUV_{mean}$ (FLAB)×TV (FLAB) (g) | | | TLG (FLAB) | 221.6±225.8 | 5.3–882.7 |

were incorporated in the multivariate analysis separately since they were found to be highly correlated (Pearson $r>0.8$, $r^2>0.66$; see the "Correlation between image-derived indices and between methodologies" section). All tests were carried out using MedCalc™ (MedCalc Software, Mariakerke, Belgium); $p$ values <0.05 were considered statistically significant.

## Results

All primary lesions were detected by $^{18}$F-FDG PET. The intensity of maximum $^{18}$F-FDG uptake in the primary lesion was quite high with a normally distributed $SUV_{max}$ of 10±3.8. As expected, $SUV_{peak}$ measurements were comparatively lower (8±3). Measurements related to the dimensions of the tumours were less uniformly distributed than SUV measurements, with a larger SD with respect to the mean. For example the TV (FLAB) was 35±33 cm$^3$.

### Correlation between image-derived indices and between methodologies

TL measurements were correlated with TV ($p<0.0001$) although with moderate coefficients ($r=0.69$, 0.58 and 0.6 for FLAB, $T_{bckgrd}^1$ and $T_{bckgrd}^2$, respectively). No significant correlation was found between any SUV measurement ($SUV_{max}$, $SUV_{peak}$, $SUV_{mean}$) and TV ($p>0.2$, $r<0.3$), irrespective of the delineation strategy, in line with results of other studies such as van Heijl et al. [33].

All $SUV_{mean}$ measurements derived from TV delineation performed using the two different methodologies considered were highly correlated ($p<0.0001$) with coefficients >0.97. TV ($r>0.82$), TL ($r>0.91$) and TLG ($r>0.95$) results were also highly correlated ($p<0.0001$) for both methodologies.

Despite high correlation coefficients, large differences were observed for several patients between measurements using the two delineation methodologies considered, and between the two users of the same adaptive thresholding. Figure 1a, b illustrates such differences. In the case of adaptive thresholding these differences were the result of the two users placing the background ROI differently.

A summary of the Bland-Altman analysis carried out to compare the delineation methods and highlight potential differences is presented in Fig. 2c, d and complete results are given in Table 3. The largest differences between methodologies were observed for TV with a bias of up to 50% between the adaptative thresholding and FLAB: both users yielded globally smaller volumes (bias of −50±50% and −21±54% for $T_{bckgrd}^1$ and $T_{bckgrd}^2$, respectively). Agreement intervals (upper and lower limits) were large for all parameters and for all comparisons, up to +80 and −140% (see Fig. 2b). Even between the two users of the same adaptive thresholding method (see Fig. 2a), mean differences of −30±35% were seen and limits of agreement were large, from −100 to +45%. No significant trend was found regarding the correlation between TV and differences between methodologies ($r<0.2$, $p>0.1$).

Better agreement was observed for TL and $SUV_{mean}$; however, intervals of agreement were large (−50 to −25% lower limit and +20 to +40% upper limit for TL; −80 to −10% lower limit and +10 to +80% upper limit for $SUV_{mean}$).

Due to the combined effect of TV and $SUV_{mean}$, TLG differences were in between, with moderate bias but still

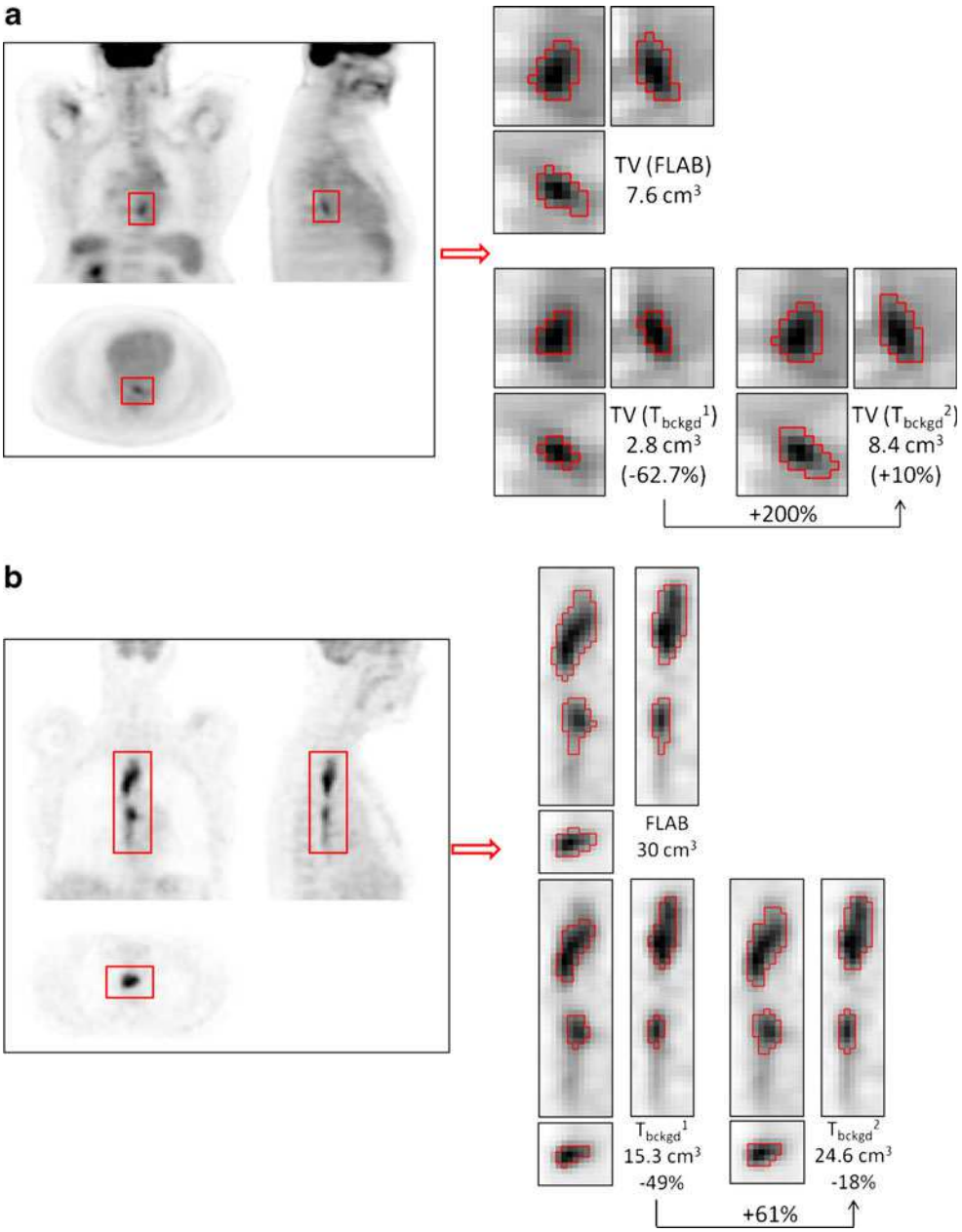



**Fig. 1** Illustration of differences in tumour delineation depending on the methodology for **a** a small (<8 $cm^3$) and low contrast (approximately 2:1) tumour and **b** a larger (30 $cm^3$) and higher contrast (approximately 7:1) tumour

large agreement intervals (upper and lower limits of −120 to −75% and +40 to +90%, respectively).

## Survival analysis

At the time of last follow-up, 10 patients were alive with no evidence of disease, 9 were alive with recurrent oesophageal cancer and 26 had died from the disease. With a median follow-up of 60 months (range 9–82), the overall median survival was 15 months. The 1-year and 2-year survival rates were 63 and 34%, respectively.

The results of the log-rank analysis of significant parameters for overall survival in univariate analysis are given in Table 4. Table 5 summarizes the prognostic value of all the parameters under investigation in this study.

Age, gender and histology types were not significant prognostic factors in the univariate analysis. Neither were T and N classifications. In the univariate analysis, the presence of metastases [median survival of 26 months (M0) versus 12 months (M1), $p=0.01$)] and the clinical AJCC stage ($p<0.001$) were significant prognostic factors.

Although there was a trend observed, neither $SUV_{max}$ nor $SUV_{peak}$ were significant prognostic factors. A $SUV_{max}$ <5 or <8 tended to be a factor for better outcome with a median survival of 14 vs 7 months ($p=0.08$) or 21 vs 13 months ($p=0.1$), respectively (see Fig. 3a).

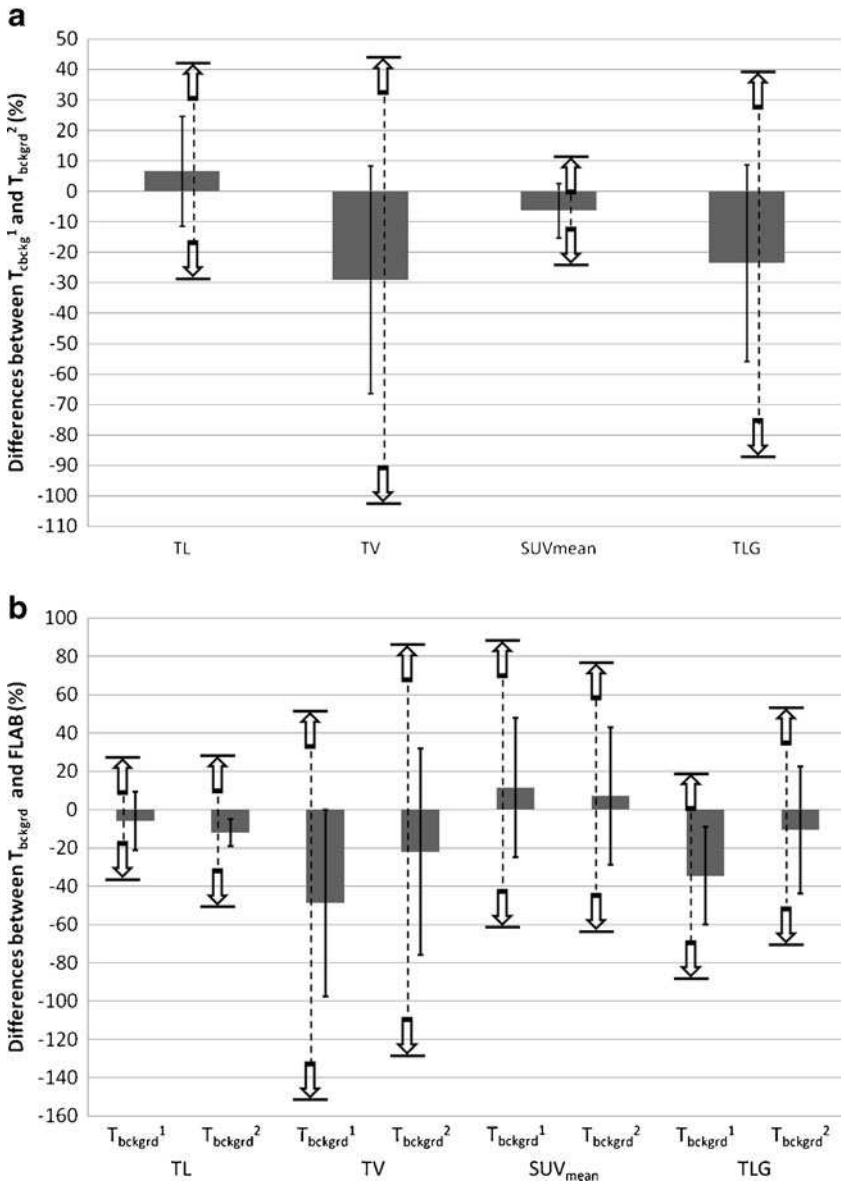



**Fig. 2** Bland-Altman analysis of differences between **a** $T_{bckgrd}^{1}$ and $T_{bckgrd}^{2}$ and **b** $T_{bckgrd}$ and FLAB, for each parameter (TL, TV, $SUV_{mean}$, TLG). *Grey columns* and *error bars* represent the mean differences (bias) and associated SD, respectively. *Bold arrows* up and down denote upper and lower limits, respectively; 95% confidence intervals for each are given in Table 3

Mean SUVs in the tumour were not significant prognostic factors in univariate analysis. There was however a trend for shorter survival associated with higher $SUV_{mean}$. For example, the median survival was reduced by a factor of 2 for patients with an $SUV_{mean}$ higher than 5 (13 months vs 21 months, $p=0.06$). This was however observed only when the FLAB methodology was used to define TV, while no similar trend was observed with $SUV_{mean}$ parameters obtained with adaptive thresholding.

Functional TV was a significant prognostic factor for overall survival, whatever methodology was used ($p<0.001$ using FLAB and $p=0.004$ for both $T_{bckgrd}^{1}$ and $T_{bckgrd}^{2}$, see Fig. 3b, c). In addition, using the TV, and independently of the delineation approach used, allowed us to separate our population into three groups with significantly different outcome ($p=0.002$, $p=0.02$ and $p=0.004$ for FLAB, $T_{bckgrd}^{1}$ and $T_{bckgrd}^{2}$, respectively). For instance, volumes defined by FLAB less than 14 $cm^3$, between 14 and 85 $cm^3$ or superior to 85 $cm^3$ were respectively associated with a median survival of 49 (19 patients), 15 (21 patients) and 5.5 (6 patients) months as illustrated in Fig. 3d. In Fig. 4a–c three examples of $^{18}$F-FDG PET baseline images of patients belonging to each of these three groups are presented.

Functional TL was also a significant prognostic factor with results similar to TV ($p=0.01$, $p=0.02$ and $p=0.04$ for FLAB, $T_{bckgrd}^{1}$ and $T_{bckgrd}^{2}$, respectively), apart from not being possible to significantly differentiate three groups of patients with different outcome, independently of the strategy.

Similarly, TLG was also a significant prognostic factor whatever methodology was used, while it was equally not possible to significantly differentiate three groups. The

**Table 3** Bland-Altman analysis results comparing delineation strategies for all parameters

| Parameter | | Mean±SD | 95% CI of mean | LL | 95% CI of LL | UL | 95% CI of UL |
|---|---|---|---|---|---|---|---|
| | | % difference between $T_{bckgrd}^1$ and $T_{bckgrd}^2$ | | | | | |
| TL | | 6.7±18 | 1.3 to 12.1 | −28.6 | −37.9 to −19.3 | 41.9 | 32.6 to 51.2 |
| TV | | −29±37.3 | −40.2 to −17.8 | −102 | −121.3 to −82.8 | 44.1 | 24.8 to 63.4 |
| $SUV_{mean}$ | | −6.3±9 | −9 to −3.6 | −23.9 | −28.5 to −19.3 | 11.2 | 6.6 to 15.8 |
| TLG | | −23.5±32.3 | −33.2 to −13.8 | −86.8 | −103.5 to −70.1 | 39.7 | 23 to 56.4 |
| | | % difference between $T_{bckgrd}$ and FLAB | | | | | |
| TL | $T_{bckgrd}^1$ | −5.9±15.3 | −10.4 to −1.4 | −35.8 | −43.6 to −28 | 24 | 16.2 to 31.8 |
| | $T_{bckgrd}^2$ | −12±7 | −18.3 to 7.1 | −49.4 | −59 to −39.9 | 24.1 | 14.5 to 33.6 |
| TV | $T_{bckgrd}^1$ | −48.8±48.8 | −63.3 to −34.3 | −144.5 | −169.5 to −120 | 46.9 | 21.9 to 71.9 |
| | $T_{bckgrd}^2$ | −22±53.9 | −38.1 to −6.0 | −127.7 | −155.3 to −100 | 83.6 | 56.1 to 111.2 |
| $SUV_{mean}$ | $T_{bckgrd}^1$ | 11.5±36.2 | 0.7 to 22.2 | −59.5 | −78 to −41 | 82.4 | 63.8 to 100.9 |
| | $T_{bckgrd}^2$ | 7.1±35.8 | −3.6 to 17.7 | −63.1 | −81.4 to −44.8 | 77.2 | 58.9 to 95.5 |
| TLG | $T_{bckgrd}^1$ | −34.5±25.6 | −42 to −26.9 | −84.6 | −97.6 to −71.5 | 15.7 | 2.6 to 28.7 |
| | $T_{bckgrd}^2$ | −10.6±33.2 | −20.4 to −0.7 | −75.6 | −92.5 to −58.6 | 54.4 | 37.5 to 71.4 |

*SD* standard deviation, *CI* confidence interval, *UL* upper limit, *LL* lower limit

median overall survival was 10 months for patients with a TLG (FLAB) >180 g and increased to 21 months for patients with a TLG<180 g ($p$=0.01). Similar results were obtained with adaptive thresholding (20 versus 8 and 20 versus 10 months for $T_{bckgrd}^1$ and $T_{bckgrd}^2$, respectively).

After multivariate analysis, considering each delineation methodology separately, only TV obtained using FLAB and AJCC stage were found to be independent significant prognostic factors ($p$=0.0017 and 0.0021 for TV and AJCC, respectively), whereas only AJCC stage was an independent significant prognostic factor ($p$<0.002) when considering TV obtained by adaptive thresholding.

Similar results were obtained when replacing TV by TL, with both TL and clinical AJCC staging found to be independent significant prognostic factors in the case of FLAB ($p$=0.017 and $p$=0.042 for AJCC stage and TL, respectively), whereas in the case of adaptive thresholding only AJCC staging was an independent significant prognostic factor ($p$=0.0021).

On the other hand, in the case of TLG only the AJCC staging was an independent significant prognostic factor ($p$<0.002), whatever delineation strategy was considered.

## Discussion

An accurate staging in oesophageal cancer is essential for guiding therapy. The standard conventional modalities are endoscopic ultrasonography and computed tomography even if this combined approach suffers from several

**Table 4** Parameters with significant prognostic value after univariate analysis

| Parameter | HR | HR 95% CI | $p$ | Median survival (months) |
|---|---|---|---|---|
| AJCC stage | 0.281 | 0.090–0.881 | 0.0008 | 26 vs 8 |
| M stage | 0.402 | 0.172–0.940 | 0.01 | 26 vs 12 |
| TL ($T_{bckgrd}^1$) | 0.318 | 0.133–0.761 | 0.02 | 21 vs 10 |
| TL ($T_{bckgrd}^2$) | 0.393 | 0.164–0.939 | 0.04 | 21 vs 10 |
| TL (FLAB) | 0.163 | 0.052–0.510 | 0.01 | 21 vs 10 |
| TV ($T_{bckgrd}^1$) | 0.212 | 0.020–2.280 | 0.004 | 16 vs 5 |
| | NA | NA | 0.02 | 21 vs 10 vs 9 |
| TV ($T_{bckgrd}^2$) | 0.212 | 0.020–2.280 | 0.004 | 16 vs 5 |
| | NA | NA | 0.004 | 49 vs 14 vs 5 |
| TV (FLAB) | 0.236 | 0.050–0.909 | 0.0005 | 20 vs 5.5 |
| | NA | NA | 0.002 | 49 vs 15 vs 5.5 |
| TLG ($T_{bckgrd}^1$) | 0.217 | 0.064–0.735 | 0.007 | 20 vs 8 |
| TLG ($T_{bckgrd}^2$ ) | 0.202 | 0.063–0.645 | 0.01 | 20 vs 10 |
| TLG (FLAB) | 0.337 | 0.147–0.772 | 0.02 | 21 vs 10 |

*HR* hazard ratio, *CI* confidence interval

**Table 5** Prognostic value of all parameters

| Variable | Significant prognostic factor in univariate analysis | Significant independent prognostic factor in multivariate analysis |
|---|---|---|
| Age | No | - |
| Gender | No | - |
| Histology type | No | - |
| AJCC stage | Yes | Yes |
| T | No | - |
| N | No | - |
| M | Yes | No |
| $SUV_{max}$ | No | - |
| $SUV_{peak}$ | No | - |
| $SUV_{mean}$ ($T_{bckgrd}^{1}$) | No | - |
| $SUV_{mean}$ ($T_{bckgrd}^{2}$) | No | - |
| $SUV_{mean}$ (FLAB) | No | - |
| TL ($T_{bckgrd}^{1}$) | Yes | No |
| TL ($T_{bckgrd}^{2}$) | Yes | No |
| TL (FLAB) | Yes | Yes |
| TV ($T_{bckgrd}^{1}$) | Yes | No |
| TV ($T_{bckgrd}^{2}$) | Yes | No |
| TV (FLAB) | Yes | Yes |
| TLG ($T_{bckgrd}^{1}$) | Yes | No |
| TLG ($T_{bckgrd}^{2}$ ) | Yes | No |
| TLG (FLAB) | Yes | No |

shortcomings. $^{18}$F-FDG PET is more and more often included in the initial staging because it allows a more accurate disease assessment, especially regarding the detection of distant metastases [2–4]. Since no patient underwent surgery in our study, anatomopathology data were not available. Therefore TNM classifications and AJCC stages were determined using suboptimal conventional staging and this could explain the poor prognostic value of T or N classification in our population.

As found in our study, $^{18}$F-FDG uptake is always present in oesophageal cancer if extended at least to submucosa [34]. Some authors suggested that the intensity of $^{18}$F-FDG uptake could be related to prognosis in oesophageal cancer, based on the good correlation existing between hexokinase activity or poor differentiation and tumour uptake [35] and also because increasing $SUV_{max}$ values seem to correlate with T classification, which is part of the TNM staging [36].

In our study, SUV measurements were not significant prognostic factors for overall survival. While various cutoff values of $SUV_{max}$ tend to be associated with a poor prognosis, none led to statistically significant differentiation. Swisher et al. reported similar results in a uniform group of highly selected patients with locally advanced oesophageal cancer treated by neoadjuvant radiochemotherapy [37]. On the other hand, these results could appear in contrast with our previous report [18], where we initially reported that an $SUV_{max}$ cutoff value of 9 had an independent prognostic value of overall survival, but this difference may be explained by the different patient characteristics considered in the two studies. We previously considered [18] a daily practice population, half of which underwent curative surgery, while we included here only patients with advanced disease exclusively treated by combined radiochemotherapy.

TL established by pathological examination has been demonstrated to be an independent prognostic factor for long-term survival [21]. Some authors proposed estimating TL based on $^{18}$F-FDG PET images using different thresholds [38]. Functional TL has been studied as a predictor of response to neoadjuvant chemoradiotherapy with conflicting results [11, 22]. In a group of 69 patients with oesophageal SCC undergoing curative surgery, Choi et al. demonstrated that functional TL was an independent prognostic factor [10]. However, one may argue that functional TL is a parameter that does not reflect the real volume of the tumour but only its longitudinal extension and could be therefore considered as only a surrogate of tumour spatial extent. This argument can be supported by the data shown in this work, where only a moderate correlation ($r<0.7$) was found between TV and TL, suggesting that functional TV may be more accurate in assessing actual tumour burden. In our study we also compared the prognostic value of TL with that of TV. Both

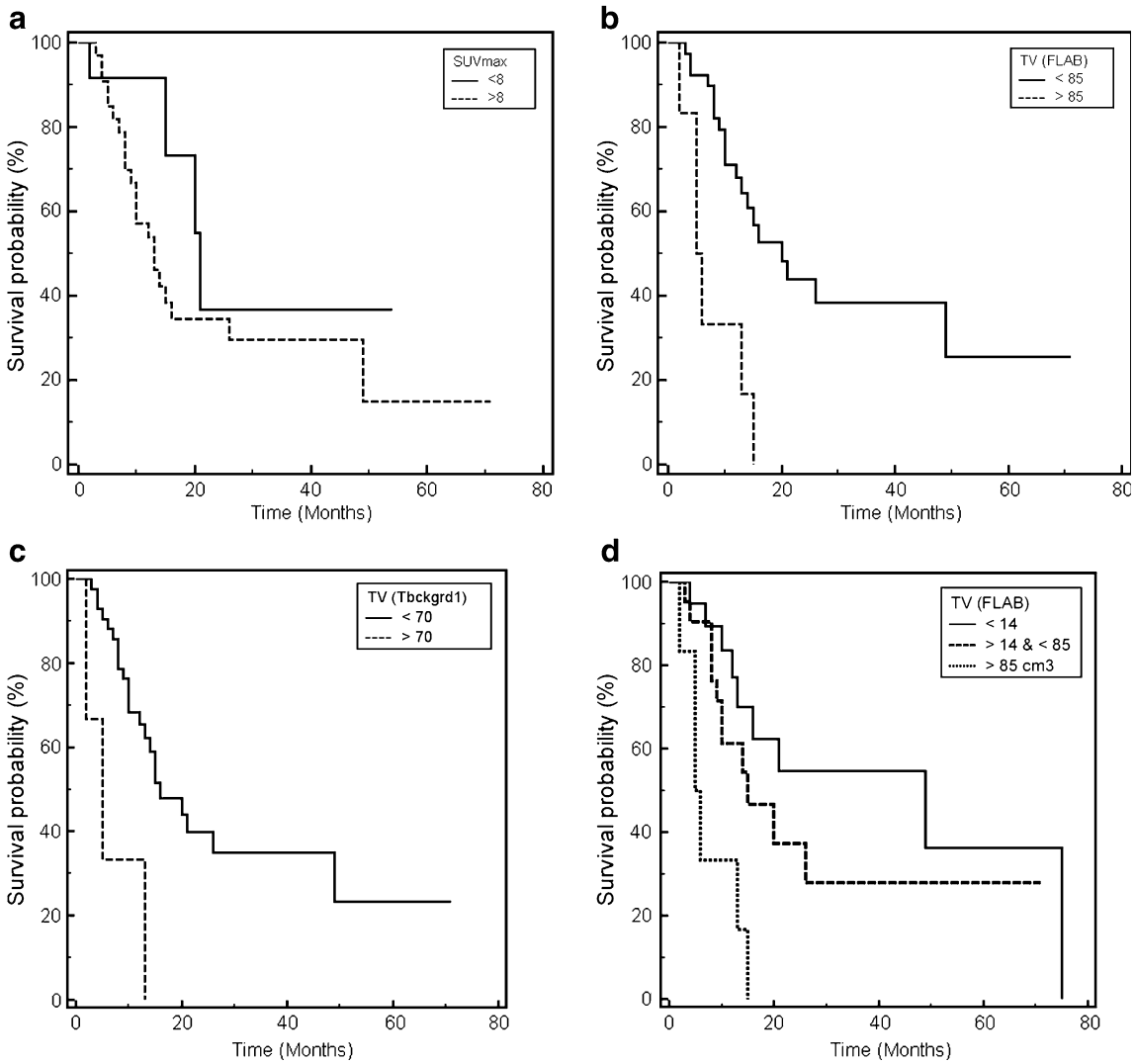



**Fig. 3** Kaplan-Meier survival curves obtained using $SUV_{max}$ (**a**), TV measured by FLAB (**b**) and $T_{bckgrd}{}^{1}$ (**c**), and defining three groups using TV measured by FLAB (**d**)

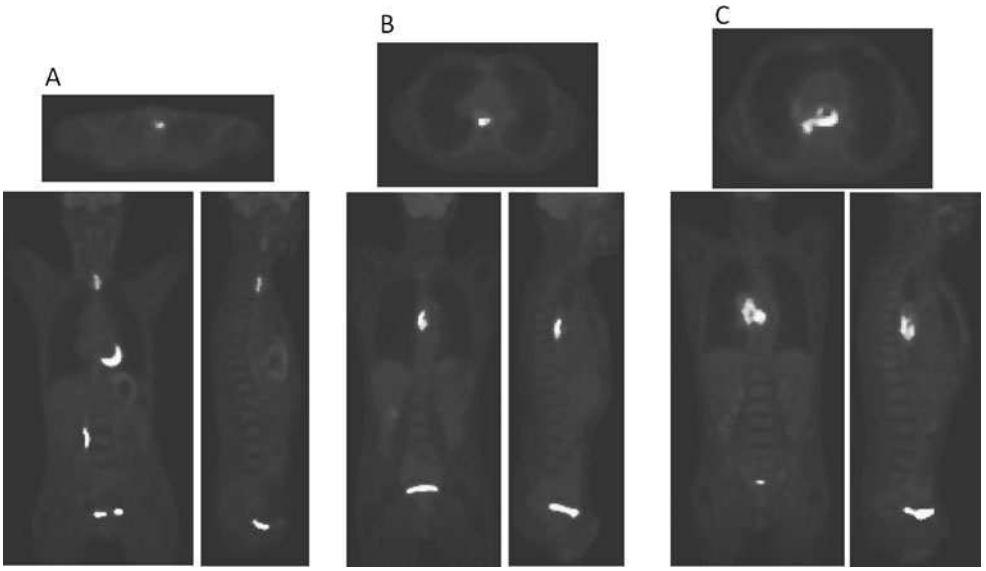



**Fig. 4** $^{18}$F-FDG PET images (axial, coronal and sagittal views from *top* to *bottom*) of patients with a **a** small tumour (11 $cm^3$, 54 months survival), **b** medium size tumour (22 $cm^3$, 18 months survival) and **c** larger tumour (92 $cm^3$, 5 months survival)

parameters were found to be significant prognostic factors irrespective of the functional volume delineation strategy. In addition, both TV and TL were independent prognostic factors for survival in the multivariate analysis. However, this result was found to be dependent on the segmentation algorithm, with both parameters being independent survival prognostic factors only when determined using the automatic FLAB segmentation. This may be related to the higher overall accuracy of FLAB with respect to adaptive thresholding for tumour delineation as previously reported [24, 27, 39]. Despite the similar prognostic values of TL and TV, only TV allowed a statistically significant stratification of patients into three groups, irrespective of the segmentation methodology. More specifically, two different cutoff values (85 and 14 $cm^3$) resulted in significant differentiation of two groups among the patients with median overall survival of 5–6 vs 20 months ($p$=0.0005) and 49 vs 13 months ($p$=0.036) for 85 and 14 $cm^3$, respectively. Being able to provide such a finer stratification of patient groups could be of value in clinical trials assessing new therapeutic regimes.

$SUV_{mean}$ measured in a volume determined using the different tumour delineation approaches considered was not found to be a prognostic factor for overall survival, although a trend was seen for $SUV_{mean}$ associated with TV defined with FLAB, which tended to differentiate patients with poor and better prognosis (13 vs 21 months, $p$=0.06).

A fundamental biological question underlying $^{18}$F-FDG PET prognostic value is whether the total volume or the metabolically active portion of the tumour is most important. Intuitively both would seem important and desirable to determine. In our study, both functional TL and TV (representative of the tumour functional spatial extent) were significant prognostic factors compared to $SUV_{mean}$ (representative of the tumour glycolytic metabolism) which was not. Combining both parameters into total lesion glycolysis index (TLG) was a prognostic factor for overall survival whatever methodology was used for tumour delineation. However, it was not an independent significant prognostic factor in the multivariate analysis. Only very few data are available on the potential clinical value of TLG in different cancer models. Xie et al. reported on the prognostic value of TLG in head and neck cancer for long-term survival [40], while Cazaentre et al. demonstrated the usefulness of TLG for predicting response to radioimmunotherapy in lymphoma [41]. To date, the limited use of TV and TLG in clinical studies could be explained by the poor accuracy, robustness and reproducibility of available tumour delineation tools affecting the clinical value of resulting measurements. The fact that TLG was not an independent prognostic factor, whereas TV alone was, suggests that the prognostic value of TLG mainly comes from the volume information and is impaired by the low prognostic value of $SUV_{mean}$ measurements. In addition, the value of TLG might be reduced by a loss of information when combining the TV and the $SUV_{mean}$ into one parameter by simple product, since large tumours with low uptake might result in the same TLG as small tumours with high uptake. Finally, the lack of partial volume effects (PVE) correction might also play a role in the reduced prognostic value of all SUV measurements as well as the resulting TLG, since tumour volumes across the patients range from quite small and significantly affected by PVE (<2 cm in diameter) to very large tumours for which PVE have smaller quantitative impact.

As expected, results concerning parameters dependent on the tumour delineation process were well correlated. On the other hand, our results also highlighted the potential impact of differences between existing tumour delineation methods, with TV and TL being independent survival prognostic factors only when determined using FLAB. This approach has been previously shown to be both robust and accurate [24, 27]. At present most commonly used methods are based on fixed or adaptive thresholds. Fixed thresholding has been demonstrated to be both inaccurate and non-robust [23, 24, 27, 39] and was therefore not considered in this study.

Regarding the adaptive thresholding performance, results from one observer ($T_{bckgrd}^2$) were closer to these of FLAB compared to the other one ($T_{bckgrd}^1$), with however significant differences, as shown in Fig. 2b and Table 3. Neither TV ($T_{bckgrd}^1$) nor TV ($T_{bckgrd}^2$) were independent prognostic factors contrary to TV (FLAB). This can be explained by the behaviour of adaptive thresholding (independently of the user) for several tumours. Most of the tumours exhibited simple shapes and homogeneous tracer uptake. However, some were more complex and exhibited higher heterogeneity, or were small (<2–3 cm) with low contrast. Adaptive thresholding has been demonstrated to provide unsatisfactory delineation for such cases [24], because its final threshold is based on the ratio between an isocontour at 70% of the maximum and the background ROI. Such an isocontour tends to overestimate (respectively underestimate) the actual value of the entire tumour for heterogeneous uptake (respectively small tumours will low contrast).

Hence the adaptive thresholding led to significant underevaluation of larger heterogeneous tumours in our study, e.g. a patient with a survival of 6 months had a TV defined by FLAB of almost 97 $cm^3$, whereas TV ($T_{bckgrd}^1$) and TV ($T_{bckgrd}^2$) were 38 $cm^3$ (−61%) and 50 $cm^3$ (−50%), respectively, clearly missing parts of the tumour. On the other hand, the dependency on the background ROI is higher regarding small tumours with low contrast, e.g. for a patient with 21 months survival, TV (FLAB) was 5.8 $cm^3$,

whereas TV ($T_{bckgrd}^1$) and TV ($T_{bckgrd}^2$) were 1.9 $cm^3$ (−67%) and 26.9 $cm^3$ (+364%), respectively. Several patients were therefore incorporated in the wrong survival curve, mostly patients with large volumes that were underestimated by the adaptive threshold.

In addition, adaptive thresholding was found to be highly user dependent, since we observed a bias up to 30% between the two users measuring TV, the agreement interval being too large for clinical applications (−110 to +45%). This seems to be in agreement with results concerning the level of reproducibility in measuring functional TV from $^{18}$F-FDG imaging which can vary from 21 to 90% using automatic and threshold-based approaches, respectively [42]. If advanced segmentation algorithms are not available, the use of adaptive thresholding approaches should be preferred to manual or fixed threshold-based delineation. Automated background ROI determination could reduce the interobserver variability observed in this work.

The limits of this study are its retrospective nature and the limited number of patients. Our results need to be confirmed through a prospective study on a larger cohort of patients. It is finally worth noting that overall survival might have been affected by other factors such as subsequent treatment for patients who relapsed, although this should have minor impact on the results of this study since it applies to all parameters considered. Other outcome measures such as progression-free survival were not investigated in this study.

## Conclusion

Our results suggest that the functional tumour volume followed by length has additional value compared to commonly used SUV measurements ($SUV_{max}$, $SUV_{peak}$, $SUV_{mean}$) for prognosis in patients with locally advanced oesophageal cancer treated with exclusive concomitant radiochemotherapy. Both parameters were significant prognostic factors for overall survival, independently of the approach used to delineate the tumours. However, only the automatic FLAB algorithm allowed TV and TL to be independent prognostic factors for survival in a multivariate analysis that included standard tumour staging. In addition, the total lesion glycolysis index was a statistically significant, but not independent, prognostic factor irrespective of the delineation algorithm used. Our findings confirm the potential value of $^{18}$F-FDG PET to give a useful orientation for patient management purposes in oesophageal cancer, but they also highlight the influence of the methodology used on the degree of pertinence of these PET image-derived parameters of interest as their accuracy and their clinical significance increase if they are computed using more reliable and robust tumour segmentation methods.

**Conflicts of interest** None.

## References

1. Falk GW. Risk factors for esophageal cancer development. Surg Oncol Clin N Am 2009;18(3):469–85.
2. Flamen P, Lerut A, Van Cutsem E, De Wever W, Peeters M, Stroobants S, et al. Utility of positron emission tomography for the staging of patients with potentially operable esophageal carcinoma. J Clin Oncol 2000;18:3202–10.
3. Heeren PA, Jager PL, Bongaerts F, van Dullemen H, Sluiter W, Plukker JT. Detection of distant metastases in esophageal cancer with (18)F-FDG PET. J Nucl Med 2004;45:980–7.
4. van Vliet EP, Heijenbrok-Kal MH, Hunink MG, Kuipers EJ, Siersema PD. Staging investigations for oesophageal cancer: a meta-analysis. Br J Cancer 2008;98(3):547–57.
5. Kim TJ, Kim HY, Lee KW, Kim MS. Multimodality assessment of esophageal cancer: preoperative staging and monitoring of response to therapy. Radiographics 2009;29(2):403–2.
6. Chuang HH, Macapinlac HA. The evolving role of PET-CT in the management of esophageal cancer. Q J Nucl Med Mol Imaging 2009;53(2):201–9.
7. MacManus M, Nestle U, Rosenzweig KE, Carrio I, Messa C, Belohlavek O, et al. Use of PET and PET/CT for radiation therapy planning: IAEA expert report 2006–2007. Radiother Oncol 2009;91(1):85–94.
8. Grégoire V, Haustermans K, Geets X, Roeis S, Lonneux M. PET-based treatment planning in radiotherapy: a new standard? J Nucl Med 2007;48(S1):68S–77.
9. Howard A, Mehta MP, Ritter MA, Bradley KA, Tome WA, Chappell RJ et al. The value of PET/CT in gross tumor volume delineation in lung and esophagus cancer. Int J Radiat Oncol Biol Phys 2004;60(Suppl):S536–7.
10. Choi JY, Jang HY, Shim YM, Kim K, Lee KS, Lee KH, et al. 18F-FDG PET in patients with esophageal squamous cell carcinoma undergoing curative surgery: prognostic implications. J Nucl Med 2004;45(11):1843–50.
11. Mamede M, Abreu-E-Lima P, Oliva MR, Nosé V, Mamon H, Gerbaudo VH. FDG-PET/CT tumor segmentation-derived indices of metabolic activity to assess response to neoadjuvant therapy and progression-free survival in esophageal cancer: correlation with histopathology results. Am J Clin Oncol 2007;30(4):377–88.
12. Hyun SH, Choi JY, Shim YM, Kim K, Lee SJ, Cho YS, et al. Prognostic value of metabolic tumor volume measured by 18F-fluorodeoxyglucose positron emission tomography in patients with esophageal carcinoma. Ann Surg Oncol 2010;17:115–22.
13. Hong D, Lunagomez S, Kim EE, Lee JH, Bresalier RS, Swisher SG, et al. Value of baseline positron emission tomography for predicting overall survival in patient with nonmetastatic esophageal or gastroesophageal junction carcinoma. Cancer 2005;104:1620–6.
14. Blackstock AW, Farmer MR, Lovato J, Mishra G, Melin SA, Oaks T, et al. A prospective evaluation of the impact of 18-F-fluorodeoxy-D-glucose positron emission tomography staging on survival for patients with locally advanced esophageal cancer. Int J Radiat Oncol Biol Phys 2006;64:455–60.
15. van Westreenen HL, Plukker JT, Cobben DC, Verhoogt CJ, Groen H, Jager PL. Prognostic value of the standardized uptake value in esophageal cancer. AJR Am J Roentgenol 2005;185(2):436–40.
16. Omloo JM, Sloof GW, Boellaard R, Hoekstra OS, Jager PL, van Dulleman HM, et al. Importance of fluorodeoxyglucose-positron emission tomography (FDG-PET) and endoscopic ultrasonography parameters in predicting survival following surgery for esophageal cancer. Endoscopy 2008;40(6):464–71.

17. Kato H, Nakajima M, Sohda M, Tanaka N, Inose T, Miyazaki T, et al. The clinical application of (18)F-fluorodeoxyglucose positron emission tomography to predict survival in patients with operable esophageal cancer. Cancer 2009;115:3196–203.
18. Cheze-Le Rest C, Metges JP, Teyton P, Jestin-Le Tallec V, Lozac'h P, Volant A, et al. Prognostic value of initial fluorodeoxyglucose-PET in esophageal cancer: a prospective study. Nucl Med Commun 2008;29:628–35.
19. Cerfolio RJ, Bryant AS. Maximum standardized uptake values on positron emission tomography of esophageal cancer predicts stage, tumor biology, and survival. Ann Thorac Surg 2006;82:391–5.
20. Rizk N, Downey RJ, Akhurst T, Gonen M, Bains MS, Larson S, et al. Preoperative 18[F]-fluorodeoxyglucose positron emission tomography standardized uptake values predict survival after esophageal adenocarcinoma resection. Ann Thorac Surg 2006;81:1076–81.
21. Yendamuri S, Swisher SG, Correa AM, Hofstetter W, Ajani JA, Francis A, et al. Esophageal tumor length is independently associated with long-term survival. Cancer 2009;115:508–16.
22. Roedl JB, Harisinghani MG, Colen RR, Fischman AJ, Blake MA, Mathisen DJ, et al. Assessment of treatment response and recurrence in esophageal carcinoma based on tumor length and standardized uptake value on positron emission tomography-computed tomography. Ann Thorac Surg 2008;86(4):1131–8.
23. Nestle U, Kremp S, Schaefer-Schuler A, Sebastian-Welsch C, Hellwig D, Rübe C, et al. Comparison of different methods for delineation of 18F-FDG PET-positive tissue for target volume definition in radiotherapy of patients with non-small cell lung cancer. J Nucl Med 2005;46(8):1342–8.
24. Hatt M, Cheze-le Rest C, Descourt P, Dekker A, De Ruysscher D, Oellers M, et al. Accurate automatic delineation of heterogeneous functional volumes in positron emission tomography for oncology applications. Int J Radiat Oncol Biol Phys 2010;77:301–8.
25. Larson SM, Erdi Y, Akhurst T, Mazumdar M, Macpinlac HA, Finn RD, et al. Tumor treatment response based on visual and quantitative changes in global tumor glycolysis using PET-FDG imaging. The visual response score and the change in total lesion glycolysis. Clin Positron Imaging 1999;2:159–71.
26. Velasquez LM, Boellaard R, Kollia G, Hayes W, Hoekstra OS, Lammertsma AA, et al. Repeatability of 18F-FDG PET in a multicenter phase I study of patients with advanced gastrointestinal malignancies. J Nucl Med 2009;50(10):1646–54.
27. Hatt M, Cheze le Rest C, Turzo A, Roux C, Visvikis D. A fuzzy locally adaptive Bayesian segmentation approach for volume determination in PET. IEEE Trans Med Imaging 2009;28(6):881–93.
28. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;1:307–10.
29. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. J Am Stat Assoc 1958;53:457–81.
30. Metz CE. Basic principles of ROC analysis. Semin Nucl Med 1978;8(4):283–98.
31. Greene FL, Page DL, Fleming ID, et al. AJCC cancer staging manual. 6th ed. New York: Springer; 2002.
32. Cox DR. Regression models and life tables. J R Stat Soc B 1972;34(2):187–220.
33. van Heijl M, Omloo JM, van Berge Henegouwen MI, van Lanschot JJ, Sloof GW, Boellaard R. Influence of ROI definition, partial volume correction and SUV normalization on SUV-survival correlation in oesophageal cancer. Nucl Med Commun 2010;31(7):652–8.
34. Himeno S, Yasuda S, Shimada H, Tajima T, Makuuchi H. Evaluation of esophageal cancer by positron emission tomography. Jpn J Clin Oncol 2002;32:340–6.
35. Fukunaga T, Okazumi S, Koide Y, Isono K, Imazeki K. Evaluation of esophageal cancers using fluorine-18-fluorodeoxyglucose PET. J Nucl Med 1998;39:1002–7.
36. Taylor MD, Smith PW, Brix WK, Wick MR, Theodosakis N, Swenson BR, et al. Correlations between selected tumor markers and fluorodeoxyglucose maximal standardized uptake values in esophageal cancer. Eur J Cardiothorac Surg 2009;35:699–705.
37. Swisher S, Erasmus J, Maish M, Correa AM, Macapinlac H, Ajani JA, et al. 2-Fluoro-2-deoxy-D-glucose positron emission tomography imaging is predictive of pathologic response and survival after preoperative chemoradiation in patients with esophageal carcinoma. Cancer 2004;101:1776–85.
38. Zhong X, Yu J, Zhang B, Li D, Han A, Song P, et al. Using 18F-fluorodeoxyglucose positron emission tomography to estimate the length of gross tumor in patients with squamous cell carcinoma of the esophagus. Int J Radiat Oncol Biol Phys 2009;73(1):136–41.
39. Tylski P, Stute S, Grotus N, Doyeux K, Hapdey S, Gardin I, et al. Comparative assessment of methods for estimating tumor volume and standardized uptake value in (18)F-FDG PET. J Nucl Med 2010;51(2):268–76.
40. Xie P, Yue JB, Zhao HX, Sun XD, Kong L, Fu Z, et al. Prognostic value of (18)F-FDG PET-CT metabolic index for nasopharyngeal carcinoma. J Cancer Res Clin Oncol 2010;136(6):883–9.
41. Cazaentre T, Morschhauser F, Vermandel M, Betrouni N, Prangère T, Steinling M, et al. Pre-therapy 18F-FDG PET quantitative parameters help in predicting the response to radioimmunotherapy in non-Hodgkin lymphoma. Eur J Nucl Med Mol Imaging 2010;37(3):494–504.
42. Hatt M, Cheze Le Rest C, Aboagye EO, et al. Reproducibility of 18F-FDG and 3′-deoxy-3′-18F-fluorothymidine PET tumor volume measurements. J Nucl Med 2010;51(9):1368–76.

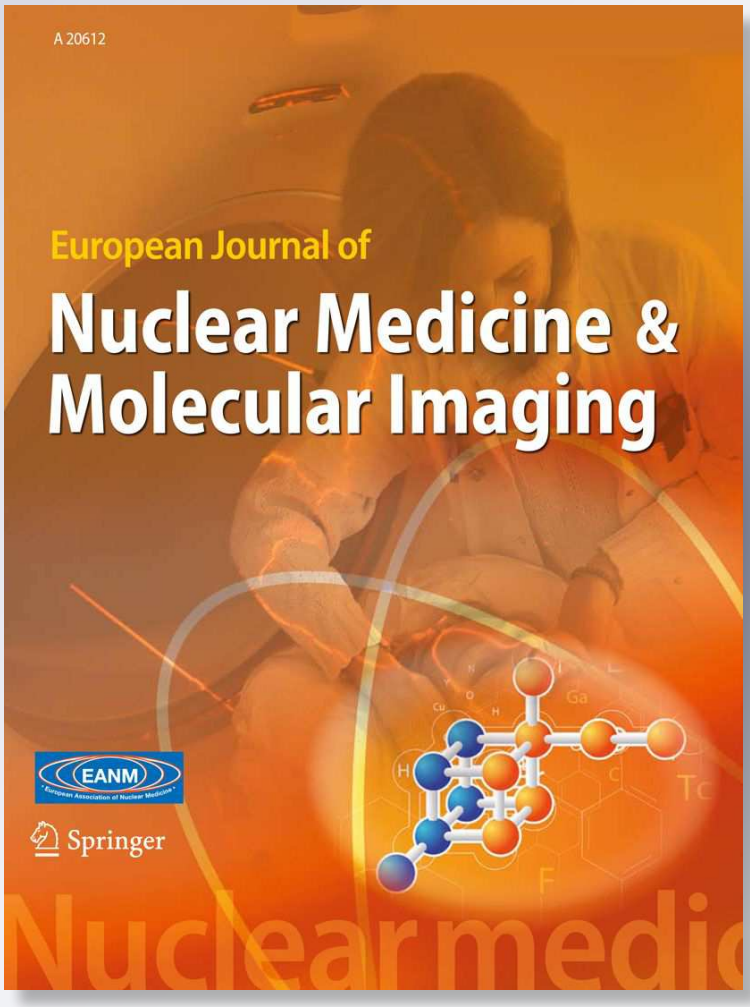
# *Baseline $^{18}$F-FDG PET image-derived parameters for therapy response prediction in oesophageal cancer*

*Mathieu Hatt, Dimitris Visvikis, Olivier Pradier & Catherine Cheze-le Rest*

**European Journal of Nuclear Medicine and Molecular Imaging**

ISSN 1619-7070
Volume 38
Number 9

Eur J Nucl Med Mol Imaging (2011) 38:1595-1606
DOI 10.1007/s00259-011-1834-9




Springer

Eur J Nucl Med Mol Imaging (2011) 38:1595–1606
DOI 10.1007/s00259-011-1834-9

ORIGINAL ARTICLE

# Baseline $^{18}$F-FDG PET image-derived parameters for therapy response prediction in oesophageal cancer

**Mathieu Hatt · Dimitris Visvikis · Olivier Pradier · Catherine Cheze-le Rest**

Received: 22 February 2011 / Accepted: 14 April 2011 / Published online: 11 May 2011


**Abstract**
*Purpose* The objectives of this study were to investigate the predictive value of tumour measurements on 2-deoxy-2-[$^{18}$F] fluoro-D-glucose ($^{18}$F-FDG) positron emission tomography (PET) pretreatment scan regarding therapy response in oesophageal cancer and to evaluate the impact of tumour delineation strategies.
*Methods* Fifty patients with oesophageal cancer treated with concomitant radiochemotherapy between 2004 and 2008 were retrospectively considered and classified as complete, partial or non-responders (including stable and progressive disease) according to Response Evaluation Criteria in Solid Tumors (RECIST). The classification of partial and complete responders was confirmed by biopsy. Tumours were delineated on the $^{18}$F-FDG pretreatment scan using an adaptive threshold and the automatic fuzzy locally adaptive Bayesian (FLAB) methodologies. Several parameters were then extracted: maximum and peak standardized uptake value (SUV), tumour longitudinal length (TL) and volume (TV), $SUV_{mean}$, and total lesion glycolysis (TLG = TV × $SUV_{mean}$). The correlation between each parameter and response was investigated using Kruskal-Wallis tests, and receiver-operating characteristic methodology was used to assess performance of the parameters to differentiate patients.
*Results* Whereas commonly used parameters such as SUV measurements were not significant predictive factors of the response, parameters related to tumour functional spatial extent (TL, TV, TLG) allowed significant differentiation of all three groups of patients, independently of the delineation strategy, and could identify complete and non-responders with sensitivity above 75% and specificity above 85%. A systematic although not statistically significant trend was observed regarding the hierarchy of the delineation methodologies and the parameters considered, with slightly higher predictive value obtained with FLAB over adaptive thresholding, and TLG over TV and TL.
*Conclusion* TLG is a promising predictive factor of concomitant radiochemotherapy response with statistically higher predictive value than SUV measurements in advanced oesophageal cancer.

**Keywords** Oesophageal cancer · Response to therapy · PET scan · Tumour volume · Total lesion glycolysis

M. Hatt (✉) · D. Visvikis · O. Pradier · C. Cheze-le Rest
LaTIM, INSERM U650, CHU Morvan,
5 avenue Foch,
29609 Brest, France
e-mail: hatt@univ-brest.fr

O. Pradier
Department of Radiotherapy, CHU Morvan,
Brest, France

## Introduction

Oesophageal cancer is the third most common malignancy of the digestive tract and a leading cause of cancer mortality worldwide with an estimated 5-year survival of 15% [1]. Despite the progress made to better understand this disease, its incidence is steadily increasing and there is a growing concern regarding its effective management [2]. The best chance for cure remains surgical resection. However, many patients have already an advanced disease (locally advanced oesophageal carcinoma, LAEC) at diagnosis and may benefit in terms of survival from neoadjuvant therapy prior to surgery [3]. The maximum benefit is for those patients who achieve a complete pathological response with no residual cancer cells in the primary tumour or lymph nodes [4]. A complete response occurs only in 15–30% of cases and is

associated with an increased overall survival [5]. On the other hand, patients who do not respond to therapy may be unnecessarily affected by toxicity of an inefficient therapy [6]. Therefore, the development of a diagnostic test offering noninvasive response to therapy prediction early in the course of treatment is of great interest, allowing potential personalization of patient management such as for inoperable tumours; chemotherapy and/or radiation therapy remains the only option. Such an assessment becomes more critical when one considers new targeted drugs that could be tested with higher efficiency if applied early [7]. For oesophageal cancer several histological markers, such as the tumour suppressor factor gene p53, the proliferative marker Ki-67 and the epidermal growth factor receptor, have been evaluated for the prediction of the therapeutic response prior to neoadjuvant therapy. None of these markers or a combination of them can currently predict response with sufficient accuracy [8, 9]. Positron emission tomography (PET) imaging with 2-deoxy-2-[$^{18}$F] fluoro-D-glucose ($^{18}$F-FDG) allows the visualization of the enhanced glucose metabolism in viable oesophageal cancer cells and may be of interest within this context. $^{18}$F-FDG PET is already well established for staging of oesophageal cancer with a better sensitivity and specificity than the combined use of CT and endoscopic ultrasonography (EUS) to detect distant metastases [10]. PET has also been shown to be promising in assessing response to therapy [11]. Several studies have shown that the reduction of the tumour's metabolic activity as measured by the standardized uptake value (SUV) from the baseline to the end of therapy uptake is predictive of a better outcome with however a large variability in the sensitivity and specificity [12]. In addition, a correlation between clinical outcome and a metabolic response observed as early as within the first 2 weeks of treatment has been demonstrated [13]. These findings suggest that tumour activity concentration differences measured on serial $^{18}$F-FDG PET scans could possibly be used to individualize treatment. However, it could be more cost-effective and beneficial to the patient to be able to predict therapy response from a single baseline PET scan acquired before the initiation of the treatment. The current study was therefore carried out to investigate the potential value of baseline $^{18}$F-FDG PET image-derived parameters for the prediction of response to combined radiochemotherapy in oesophageal cancer. A secondary objective was to investigate the potential influence of the method used to delineate the tumour on the prediction results.

## Materials and methods

### Patients

Fifty consecutive patients with newly diagnosed oesophageal cancer treated with exclusive concomitant radiochemotherapy between 2004 and 2008 were included in this study. As part of the routine procedure for the initial staging in oesophageal cancer, each patient was referred for an $^{18}$F-FDG PET study before treatment. It included three courses of 5-fluorouracil/cisplatin and a median radiation dose of 60 Gy given in 180-cGy daily fractions delivered once daily, 5 days a week for 6–7 weeks. The characteristics of the patients are given in Table 1. Most of them (45 of 50) were male, aged 65± 9 years at the time of diagnosis; 74% of the tumours, most of which were squamous cell carcinoma (72%), originated from the middle and lower oesophagus. Response to therapy was evaluated 1 month after the completion of the concomitant radiochemotherapy using conventional thoraco-abdominal

**Table 1** Patient demographic and clinical characteristics

| Parameter | Number of patients (%) |
|---|---|
| Gender | |
| Male | 45 (90) |
| Female | 5 (10) |
| Age | |
| Range | 45–84 |
| Median | 69 |
| Site | |
| Upper oesophagus | 13 (26) |
| Middle oesophagus | 20 (40) |
| Lower oesophagus | 17 (34) |
| Histology type | |
| Adenocarcinoma | 14 (28) |
| Squamous cell carcinoma | 36 (72) |
| Histology differentiation | |
| Well differentiated | 14(28) |
| Moderately differentiated | 12 (24) |
| Poorly differentiated | 5 (10) |
| Unknown | 19 (38) |
| TNM stage | |
| T1 | 7 (14) |
| T2 | 8 (16) |
| T3 | 24 (48) |
| T4 | 11 (22) |
| N0 | 20 (40) |
| N1 | 30 (60) |
| M0 | 34 (68) |
| M1 | 16 (32) |
| AJCC stage | |
| I | 4 (8) |
| IIA | 8 (16) |
| IIB | 6 (12) |
| III | 16 (32) |
| IVA | 16 (32) |

CT and endoscopy. Patients were classified as non-responders (NR) including stable and progressive disease, partial responders (PR) or complete responders (CR). Response evaluation was based on CT evolution between pretreatment and post-treatment scans using Response Evaluation Criteria in Solid Tumors (RECIST) [14]. Patients also underwent fibroscopy in cases of partial or complete response. Complete response was confirmed by the absence of visible disease in the high endoscopy and no viable tumour on biopsy. Partial CT response was confirmed by macroscopic residual (>10% viable) on biopsy. No discordance was observed between pathological, when available, and CT evaluation.

The current analysis was carried out after an approval by the Institutional Ethics Review Board.

## $^{18}$F-FDG PET acquisitions

All $^{18}$F-FDG PET studies were carried out prior to the initiation of treatment. Patients were instructed to fast for at least 6 h before the $^{18}$F-FDG administration (5 MBq/kg). Static emission images were acquired from head to thigh (2 min per bed position) beginning 60 min after injection on a Philips GEMINI PET/CT system (Philips Medical Systems, Cleveland, OH, USA). Images were reconstructed using the RAMLA 3D algorithm and CT based attenuation correction. Optimized reconstruction parameters were used for the 3-D row action maximum likelihood algorithm (RAMLA) based on the standard optimized clinical protocol (2 iterations, relaxation parameter of 0.05, 5 mm 3-D Gaussian post-filtering, 4×4×4 mm$^3$ voxels grid sampling). The PET images were corrected for attenuation using CT-based attenuation correction.

## PET image analysis

All parameters considered were extracted from the baseline PET images only. For each patient, the primary tumour was identified on the baseline pretreatment PET images by a nuclear physician. Three different SUV measurements and three parameters related to the tumour functional dimensions, namely the tumour volume (TV), tumour longitudinal length (TL) and total lesion glycolysis (TLG) [15], were extracted for each primary lesion. SUV measurements considered were $SUV_{max}$, $SUV_{peak}$ defined as the mean of $SUV_{max}$ and its 26 neighbours [roughly similar to a 1-cm region of interest (ROI)] and mean SUV within the delineated tumour ($SUV_{mean}$). Whereas $SUV_{max}$ and $SUV_{peak}$ are clearly independent of the tumour delineation strategy used, TL, TV, $SUV_{mean}$ and the derived TLG values might depend on the delineation process. To study the impact of this step, we considered two different approaches, namely the automatic fuzzy locally adaptive Bayesian (FLAB) algorithm [16, 17] and an adaptive threshold algorithm [18] optimized for the GEMINI PET/CT scanner. Although the first approach is fully automatic, adaptive thresholding requires a manually defined background ROI. Therefore, two experienced nuclear medicine physicians were considered in the background ROI definition, leading to two series of results denoted as $T_{A1}$ and $T_{A2}$. TL was determined in longitudinal direction by multiplying the number of slices in the delineated TV by the PET image slice thickness (4 mm). TV was defined as the sum of all voxels contained in the delineated volumes multiplied by the image voxel's volume (64 mm$^3$). Finally, TLG was determined by multiplying the $SUV_{mean}$ and associated TV.

**Table 2** Definition of image-derived parameters and associated summary statistics

| Definition | | | Notation | Mean±SD |
|---|---|---|---|---|
| Highest SUV | | | $SUV_{max}$ | 9.7±3.9 |
| Mean of $SUV_{max}$ and its 26 neighbours | | | $SUV_{peak}$ | 8.0±3.3 |
| Mean SUV within tumour defined by | Adaptive threshold | User 1 | $SUV_{mean}(T_{A1})$ | 6.4±2.5 |
| | | User 2 | $SUV_{mean}(T_{A2})$ | 6.0±2.6 |
| | FLAB | | $SUV_{mean}$(FLAB) | 5.5±2.3 |
| TL (cm) defined by | Adaptive threshold | User 1 | $TL(T_{A1})$ | 5.8±2.9 |
| | | User 2 | $TL(T_{A2})$ | 5.5±2.8 |
| | FLAB | | TL(FLAB) | 6.0±2.8 |
| TV (cm$^3$) defined by | Adaptive threshold | User 1 | $TV(T_{A1})$ | 27.2±25.6 |
| | | User 2 | $TV(T_{A2})$ | 34.8±30.7 |
| | FLAB | | TV(FLAB) | 39.4±34.9 |
| $SUV_{mean}(T_{A1}) \times TV(T_{A1})$(g) | | | $TLG(T_{A1})$ | 175.6±178.9 |
| $SUV_{mean}(T_{A2}) \times TV(T_{A2})$(g) | | | $TLG(T_{A2})$ | 206.9±203.4 |
| $SUV_{mean}$(FLAB) × TV(FLAB)(g) | | | TLG(FLAB) | 207.3±192.0 |

Statistical analysis

The relation between response to therapy and each parameter distribution was studied using the Kruskal-Wallis test [19] as recommended for small, not normally distributed samples.

Receiver-operating characteristic (ROC) methodology [20] was used to assess the performance of each parameter to differentiate patients. Two classification tasks were considered: differentiating CR patients from PR and NR, or NR patients from CR and PR. Evaluation was performed

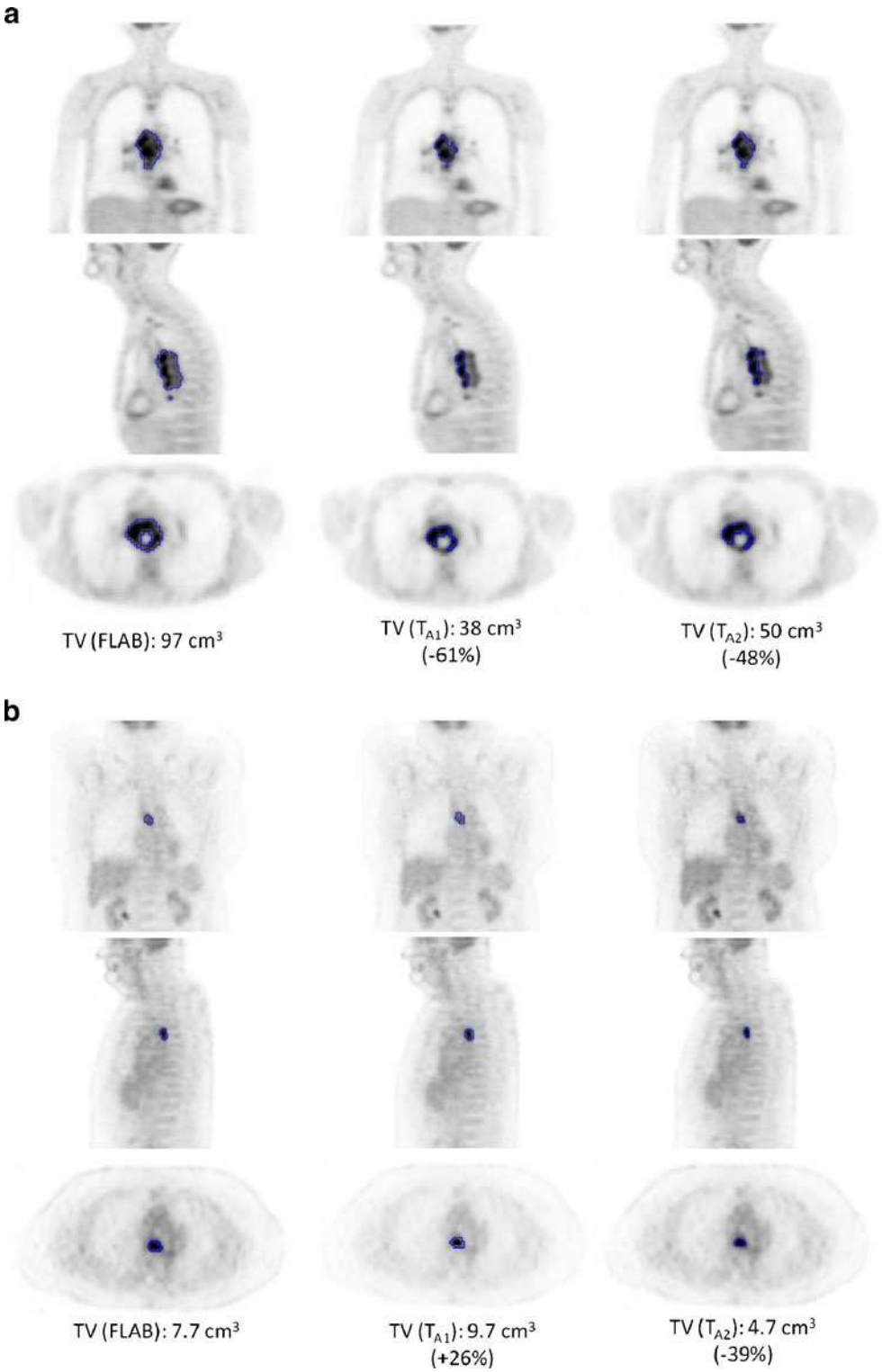



**Fig. 1** Illustration of differences in tumour delineation depending on the methodology for two patients

in terms of the area under the curve (AUC) as well as specificity and sensitivity.

The significance of the following factors was tested: age, gender, T, N and M classifications, American Joint Committee on Cancer (AJCC) stage, histology types, $SUV_{max}$, $SUV_{peak}$, TL, TV, $SUV_{mean}$ and TLG. All tests were two-sided and $p$ values <0.05 were considered statistically significant.

## Results

The range of values for the different image-derived indices as well as the mean and standard deviation for the patient population considered are given in Table 2. All primary lesions were detected by $^{18}$F-FDG PET exhibiting a rather high uptake with an $SUV_{max}$ of 9.7±3.9. $SUV_{peak}$ and $SUV_{mean}$ measurements were comparatively lower (8.0±3.3 and 5.8±2.4, respectively).

### Correlation between image-derived indices and between methodologies

TV and TL measurements were moderately correlated ($r$=0.77, 0.68 and 0.60 for FLAB, $T_{A1}$ and $T_{A2}$ respectively, $p$<0.0001). On the other hand, no significant correlation was found between TV and any of the SUV measurements ($r$<0.2, $p$>0.1), irrespective of the delineation approach used. High correlations were observed between the TV ($r$>0.89), TL ($r$>0.90) or TLG ($r$>0.93) measurements obtained with the two delineation strategies ($p$<0.0001). Even higher correlation coefficients ($r$>0.97, $p$<0.0001) were observed for the $SUV_{mean}$ measurements derived using the two different tumour segmentation approaches (FLAB and adaptive thresholding). Despite these correlations, certain large differences were observed for a few patients between the delineation results of the two segmentation algorithms considered, examples of which are illustrated in Fig. 1.

### Response to therapy analysis

Of the 50 patients included in the study, 25 were classified as PR, while there were 12 CR and 13 NR. Results concerning the predictive value of all parameters considered are summarized in Tables 3 and 4 containing the results of the Kruskal-Wallis tests and that of the ROC analysis (considering the AUC, specificity and sensitivity regarding the classification tasks), respectively.

**Table 3** Kruskal-Wallis test results for each parameter considering the ability to differentiate ($p$<0.05) each pair of response group

| Kruskal-Wallis tests | | | | | | |
|---|---|---|---|---|---|---|
| Parameter | | Test statistic | $p$ | Response differentiation? ($p$<0.05) | | |
| | | | | CR ($n$=12) vs NR ($n$=13) | CR ($n$=12) vs PR ($n$=25) | PR ($n$=25) vs NR ($n$=13) |
| Age | | 0.4 | 0.83 | No | No | No |
| Gender | | 4.0 | 0.14 | No | No | No |
| T | | 4.9 | 0.09 | No | No | No |
| N | | 2.7 | 0.26 | No | No | No |
| M | | 3.6 | 0.17 | No | No | No |
| AJCC stage | | 5.9 | 0.052 | Yes | No | Yes |
| Histology type | | 2.3 | 0.32 | No | No | No |
| $SUV_{max}$ | | 2.5 | 0.29 | No | No | No |
| $SUV_{peak}$ | | 3.9 | 0.14 | No | No | No |
| $SUV_{mean}$ | $T_{A1}$ | 3.3 | 0.197 | No | No | No |
| | $T_{A2}$ | 3.2 | 0.199 | No | No | No |
| | FLAB | 2.6 | 0.270 | No | No | No |
| TL | $T_{A1}$ | 14.5 | 0.0007 | Yes | Yes | Yes |
| | $T_{A2}$ | 12.4 | 0.0020 | Yes | Yes | Yes |
| | FLAB | 15.6 | 0.0004 | Yes | Yes | Yes |
| TV | $T_{A1}$ | 13.9 | 0.0010 | Yes | Yes | Yes |
| | $T_{A2}$ | 12.9 | 0.0016 | Yes | Yes | Yes |
| | FLAB | 16.2 | 0.0003 | Yes | Yes | Yes |
| TLG | $T_{A1}$ | 14.6 | 0.0007 | Yes | Yes | Yes |
| | $T_{A2}$ | 14.6 | 0.0007 | Yes | Yes | Yes |
| | FLAB | 21.1 | <0.0001 | Yes | Yes | Yes |

**Table 4** ROC analysis results with AUC and associated 95% confidence intervals (CI), specificity (Sp) and sensitivity (Se) for each parameter regarding the two classification tasks

| Parameter | | ROC analysis | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | NR&PR ($n$=38) vs CR ($n$=12) | | | | NR ($n$=13) vs PR&CR ($n$=37) | | | |
| | | AUC | 95% CI | Se (%) | Sp (%) | AUC | 95% CI | Se (%) | Sp (%) |
| Age | | 0.51 | 0.32–0.70 | 83.3 | 31.6 | 0.55 | 0.35–0.75 | 86.5 | 36.5 |
| Gender | | 0.61 | 0.46–0.75] | 27.3 | 94.3 | 0.51 | 0.41–0.62 | 90.9 | 11.4 |
| T | | 0.70 | 0.47–0.93 | 60.0 | 89.3 | 0.64 | 0.49–0.78 | 100.0 | 33.3 |
| N | | 0.64 | 0.46–0.83 | 60.0 | 68.8 | 0.55 | 0.38–0.73 | 70.0 | 40.6 |
| M | | 0.56 | 0.38–0.73 | 70.0 | 41.2 | 0.70 | 0.53–0.87 | 70.0 | 70.6 |
| AJCC stage | | 0.63 | 0.43–0.84 | 54.6 | 73.5 | 0.72 | 0.57–0.88 | 87.5 | 46.2 |
| Histology type | | 0.51 | 0.35–0.66 | 72.7 | 28.2 | 0.60 | 0.46–0.75 | 42.9 | 77.8 |
| $SUV_{max}$ | | 0.65 | 0.45–0.85 | 33.3 | 94.7 | 0.54 | 0.34–0.73 | 30.8 | 89.2 |
| $SUV_{peak}$ | | 0.69 | 0.49–0.89 | 75.0 | 63.2 | 0.54 | 0.35–0.73 | 30.8 | 86.5 |
| $SUV_{mean}$ | $T_{A1}$ | 0.67 | 0.47–0.87 | 50.0 | 84.2 | 0.54 | 0.35–0.74 | 89.2 | 38.0 |
| | $T_{A2}$ | 0.67 | 0.45–0.88 | 50.0 | 94.7 | 0.51 | 0.32–0.70 | 100.0 | 16.2 |
| | FLAB | 0.65 | 0.43–0.87 | 58.3 | 84.2 | 0.51 | 0.32–0.70 | 100.0 | 13.5 |
| TL | $T_{A1}$ | 0.81 | 0.65–0.97 | 83.3 | 79.0 | 0.78 | 0.63–0.93 | 59.5 | 92.3 |
| | $T_{A2}$ | 0.79 | 0.63–0.96 | 83.3 | 73.3 | 0.75 | 0.61–0.90 | 75.7 | 69.2 |
| | FLAB | 0.79 | 0.64–0.94 | 83.3 | 65.8 | 0.82 | 0.70–0.94 | 59.5 | 92.3 |
| TV | $T_{A1}$ | 0.79 | 0.65–0.89 | 75.0 | 81.6 | 0.79 | 0.65–0.93 | 78.4 | 69.2 |
| | $T_{A2}$ | 0.74 | 0.59–0.85 | 83.3 | 57.9 | 0.81 | 0.67–0.95 | 94.6 | 53.9 |
| | FLAB | 0.78 | 0.64–0.88 | 75.0 | 79.0 | 0.84 | 0.72–0.96 | 75.7 | 76.9 |
| TLG | $T_{A1}$ | 0.81 | 0.62–1.00 | 66.7 | 92.1 | 0.78 | 0.65–0.92 | 92.3 | 56.8 |
| | $T_{A2}$ | 0.80 | 0.61–0.99 | 75.0 | 86.8 | 0.80 | 0.67–0.93 | 69.2 | 81.1 |
| | FLAB | 0.85 | 0.73–0.98 | 75.0 | 92.1 | 0.86 | 0.75–0.98 | 84.6 | 75.7 |

Age, gender or T, N and M classifications did not allow significant prediction of the response to treatment. The AJCC stage was not significantly ($p$>0.05) associated with the type of response, despite the fact that all NR were at least stage IIB and could be statistically differentiated from both PR and CR ($p$<0.05). However, AJCC stage could not differentiate PR from CR ($p$>0.05). Finally, there was no statistical correlation between histology type and response ($p$=0.3).

Figure 2 shows a graphical comparison of the Kruskal-Wallis results considering the predictive value of the different SUV parameters considered. Initial $SUV_{max}$ (Fig. 2a) was not predictive of response to therapy ($p$=0.29) although CR tended to have smaller $SUV_{max}$ (8.1±4.1) than PR and NR (10.2±3.7 and 10.2±3.9, respectively). Similarly, $SUV_{peak}$ (Fig. 2b) was not predictive of response to therapy with a mean value of 6.5±3.5 in CR, whereas both PR and NR were characterized by similar higher $SUV_{peak}$ values (8.5±3.1 and 8.4±3.3, respectively) ($p$=0.14). None of the $SUV_{mean}$ measurements, whatever delineation strategy was used, could significantly predict response to therapy ($p$>0.19).

On the contrary, all parameters related to tumour spatial extent (TL, TV and TLG) measurements allowed significant ($p$<0.002) differentiation of the three response groups, irrespective of the segmentation methodology (see Fig. 3a–c). For instance, TV as measured by FLAB was 20±25, 32±24 and 72±40 $cm^3$ for CR, PR and NR patients, respectively. The parameter that allowed the best differentiation between the three patient groups was TLG measured by FLAB (Kruskal-Wallis test $p$<0.0001, see Fig. 3c), with a TLG of 74±75, 179±143 and 385±226 g for CR, PR and NR patients, respectively. Figure 4 shows examples of one CR, one PR and one NR patient with corresponding TLG values.

The ROC analysis results confirmed the limited predictive value of most SUV measurements for the accurate classification of either CR vs PR and NR, or NR vs PR and CR (AUC<0.70 and <0.56, respectively). Differences between ROC analysis associated with SUV measurements and those associated with TL, TV and TLG were significant ($p$<0.05) for both tasks (see examples in Fig. 5). Better predictive performances were obtained with TL, TV and TLG measurements with significantly higher AUC (from 0.74 to 0.86) for both tasks ($p$<0.05). For instance, using FLAB a TLG <58 g allowed identifying CRs with a sensitivity of 75% and a specificity of 92%, and a TLG >196 g identified NRs with a sensitivity of 76% and a specificity of 85%. However, in terms of predictive performance no significant differences were obtained between TL, TV and TLG measurements for both tasks.

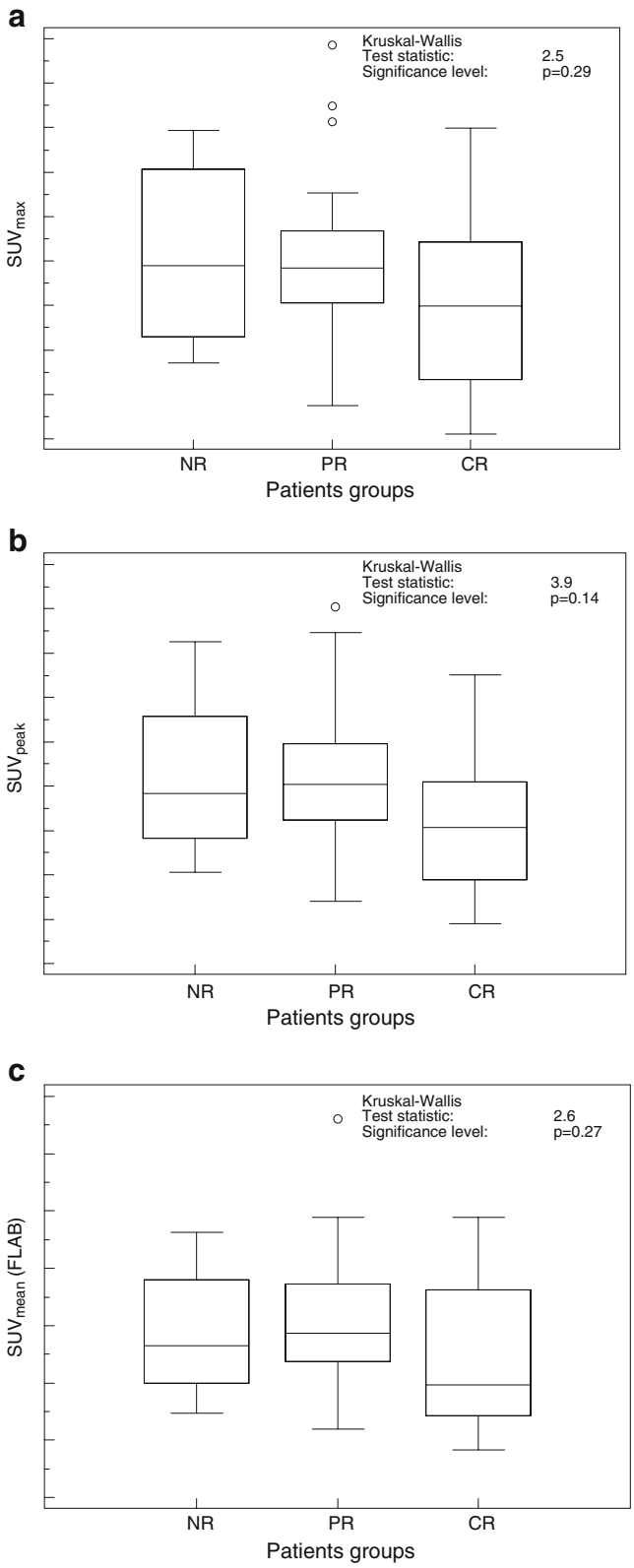



**Fig. 2** Distributions of NR, PR and CR patients and associated Kruskal-Wallis tests for SUV-based image-derived indices: $SUV_{max}$ (**a**) and $SUV_{peak}$ (**b**)

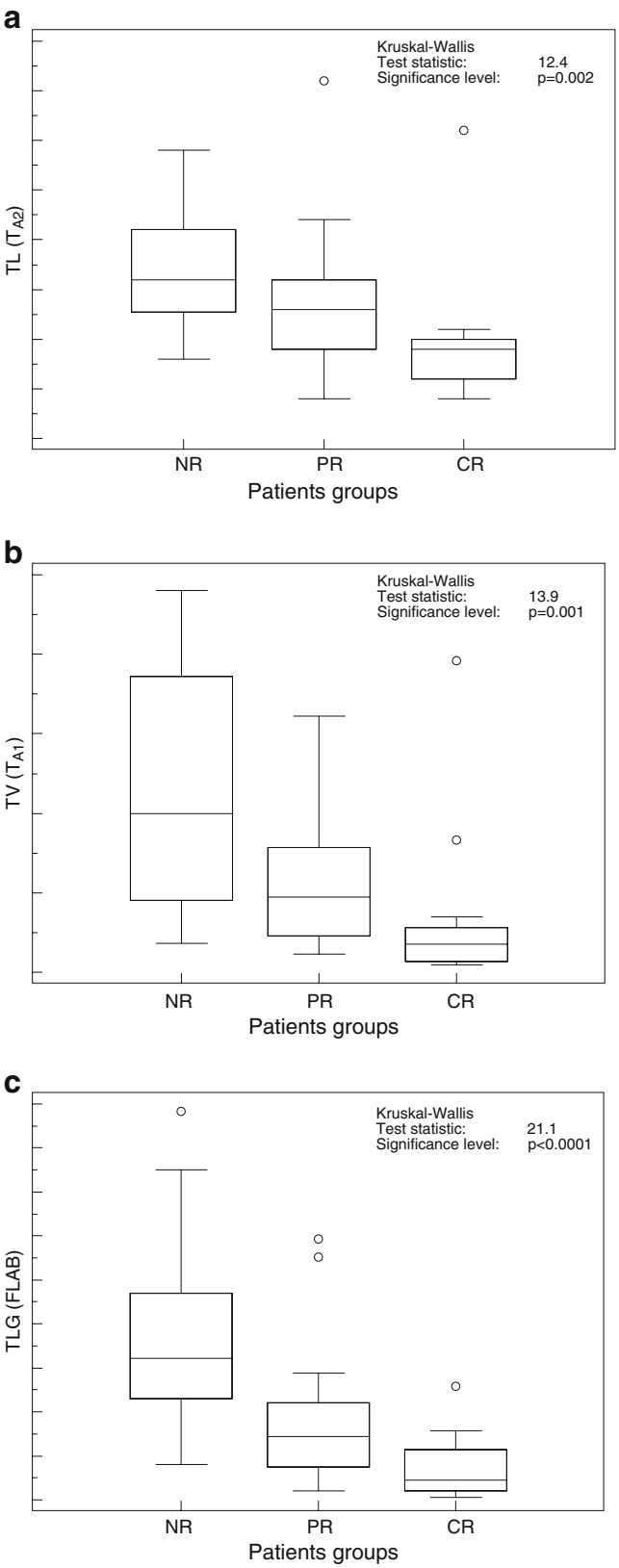



**Fig. 3** Distributions of NR, PR and CR patients and associated Kruskal-Wallis tests for TV-related image-derived indices: TL ($T_{A2}$) (**a**), TV ($T_{A1}$) (**b**) and TLG (FLAB) (**c**)

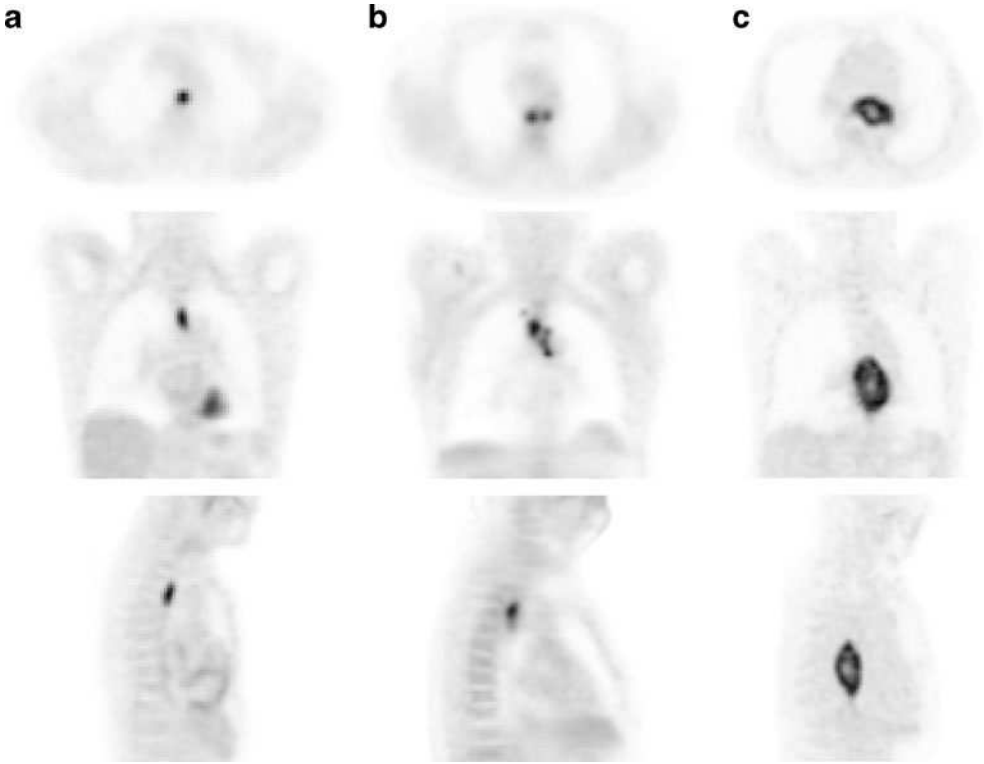



**Fig. 4** $^{18}$F-FDG PET axial, coronal and sagittal images of a complete responder with 20 g TLG (**a**), partial responder with 100 g TLG (**b**) and non-responder with 750 g TLG (**c**)

In terms of an observed trend, better results were obtained for TLG over TV and TL whatever tumour delineation approach was used (Tables 3 and 4). In addition, there was a systematic although not statistically significant trend of better performance for those parameters when obtained with FLAB compared to the use of the adaptive threshold, as demonstrated by higher AUC and smaller confidence intervals, as well as higher sensitivity and specificity for both classification tasks (Table 4).

The analysis with respect to histology type (adenocarcinoma vs squamous cell carcinoma) led to results similar to what was observed when considering the entire population. Within the same context no statistically significant differences were observed between the two patient groups in the hierarchy of parameters and results derived using the different functional TV delineation methods.

The predictive value of TLG, combining TV and $SUV_{mean}$ into one single parameter, was higher than the one of TV, despite the non-significant value of $SUV_{mean}$ alone. Considering together TV and $SUV_{mean}$, one is able to differentiate different treatment response patient groups (see Fig. 6). On the one hand, TLG increased the differentiation between CR and NR, as all NR had either a TV above 50 cm$^3$ (8/13) or an $SUV_{mean}$ above 5 (8/13), while 10 of 12 CR had either a small TV (<15 cm$^3$) (9/12) or $SUV_{mean}$ (<5) (7/12), and half of them (6/12) had both. On the other hand, PR had either a higher $SUV_{mean}$ than CR for volumes below 25 cm$^3$ (6.5±2.7 vs 4.5±2.4) or lower $SUV_{mean}$ than NR for TV of 25–50 cm$^3$ (5.8±1.8 vs 7.1±0.9). Therefore, the use of TLG increased the differentiation between PR and CR, as well as between PR and NR for volumes below 15 cm$^3$ and between 25 and 50 cm$^3$, respectively.

## Discussion

Assessment of response to therapy early during treatment plays an important role in patient management as well as in drug development and new criteria including PET have been suggested for this task [21, 22]. However, being able to predict response to therapy before the initiation of the treatment would be even more powerful for patient management. In this context, either patient or tumour characteristics could be considered. In our study we focused on functional imaging and different image-derived parameters related to tumour uptake using PET. The results of our study demonstrate that TV-based parameters derived from baseline $^{18}$F-FDG PET images in oesophageal cancer are good predictors of response to therapy, with high TL, TV and TLG being associated with poor response to combined radiochemotherapy. On the contrary, more commonly used parameters such as tumour SUVs were not predictors of response to therapy considering only the baseline $^{18}$F-FDG PET images. These results further demonstrate the value of TV-based PET image-derived parameters, since we have previously demonstrated a superior prognostic value of baseline functional TL, TV and TLG over SUV measurements for

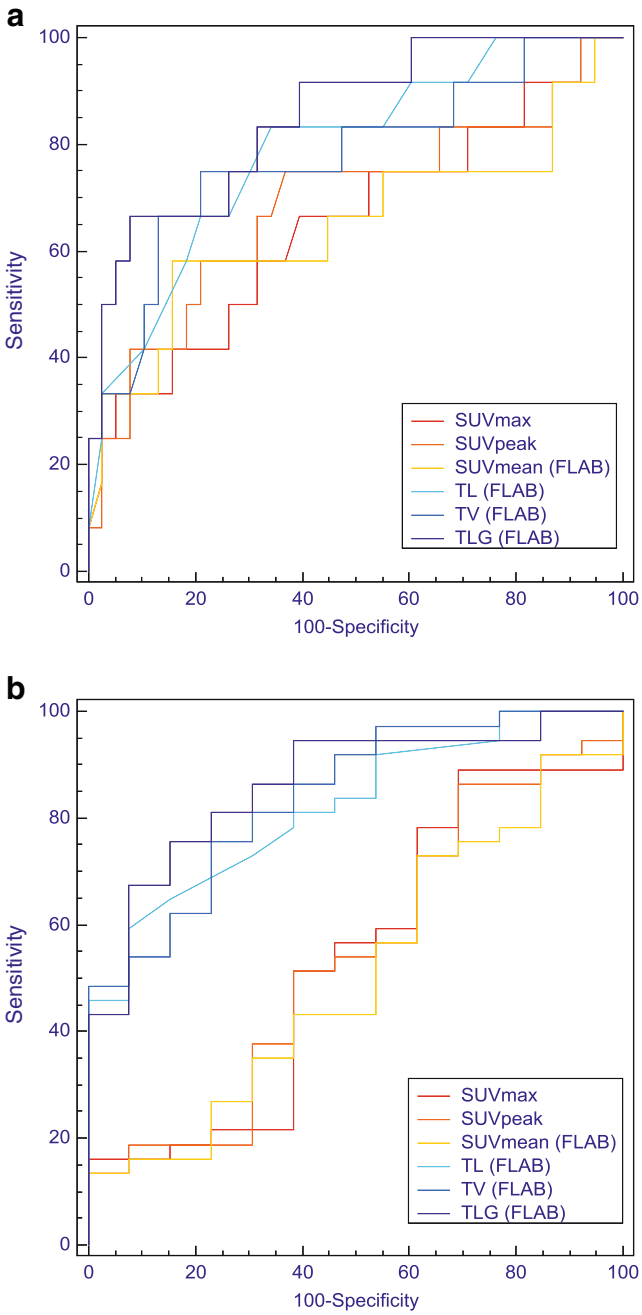



**Fig. 5** Examples of ROC curves obtained for classification tasks of differentiating CR from NR&PR (**a**) or NR from PR&CR (**b**). Comparison of ROC curves for SUV measurements ($SUV_{max}$ in *red*, $SUV_{peak}$ in *orange* and $SUV_{mean}$ in *yellow*) and TL, TV and TLG measured with FLAB (in *light blue*, *blue* and *dark blue*, respectively)

overall survival in a similar group of oesophageal cancer patients [23].

FDG PET has been previously used for the prediction of response to therapy or prognosis in a variety of malignancies [24]. Considering the predictive value of baseline FDG uptake for therapy response in oesophageal cancer, only few data showing conflicting results are available [12]. Levine et al. and Rizk et al. reported a high initial $SUV_{max}$ being associated with good response [25, 26], whereas Makino et al. and Kato et al. found the opposite [27, 28]. These conflicting results can be potentially attributed to differences in patient populations, tumour histology types, as well as treatment, but could also suggest that SUV measurements are unreliable in this context. Although similarly to the results of Kato et al. and Makino et al. our results suggest that lower values of $SUV_{max}$ are associated with a complete response, this trend was not significant. In addition, $SUV_{mean}$ or $SUV_{peak}$, considered more robust to potential noise bias associated with $SUV_{max}$, were also not significant predictors of response to therapy in our study.

One of the demonstrated independent predictors of long-term survival in oesophageal cancer is longitudinal tumour extension established by pathological examination [29]. It has been previously demonstrated that TL measured on CT images leads to a weak correlation with the pathological TL, associated with a large overestimation [30]. Some authors proposed the estimation of metabolic TL as a surrogate of pathological TL using various thresholds of $^{18}$F-FDG PET uptake [31]; however, conflicting results concerning the predictive value of metabolic TL for response to neoadjuvant radiochemotherapy have been observed [32, 33]. One may argue that TL does not reflect the entire volume of the tumour and could therefore be only considered as a limited surrogate measure of tumour functional spatial extent. This assumption is partly supported by our data, in which only a moderate correlation ($r$ between 0.6 and 0.77) was found between TV and TL, suggesting that TV may bring additional information compared to TL in assessing overall tumour burden. In our study both TV and TL were found to be significant predictive factors of response to therapy, irrespective of the functional volume delineation strategy, with only a small and non-significant improvement of the predictive value of TV over TL.

TV and TLG measured on PET are 3-D measurements incorporating metabolically active TV not available from CT data [34]. It has already been demonstrated that a decrease of the TV and TLG can predict response to therapy [35, 36]. These studies however have explored differences in indices derived from serial PET images. The value of such indices obtained on the baseline scan only within the context of therapy response prediction in oesophageal cancer has not previously been explored. Because these parameters reflect metabolic information in the entire tumour, they may be more accurate for tumour characterization than a single voxel measure and this may explain why TV and TLG were good predictors of therapy response as demonstrated in our study. Our results are consistent with recent studies in pleural mesothelioma and lymphoma patients that have demonstrated the potential of such

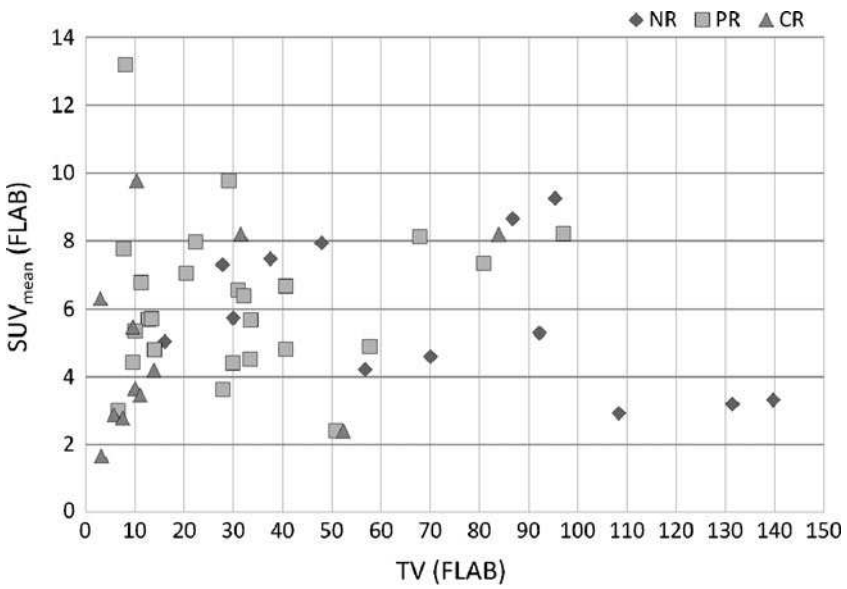



**Fig. 6** Distribution of CR, PR and NR patients according to their $SUV_{mean}$ and TV as measured by FLAB

indices extracted from baseline $^{18}$F-FDG PET scan to predict response to therapy [37, 38].

Despite a great potential value, such indices have been only of limited use to date, which can be explained by the limited accuracy, robustness and reproducibility of the available tumour delineation tools [39, 40]. In oesophageal cancer only the prognostic value of TV has been studied [23, 41], while there are limited data on the value of TLG [23]. In our study TLG allowed identifying complete responders and non-responders with moderate sensitivity (75 and 76%, respectively) and high specificity (92 and 85%, respectively). Prospective studies with a larger patient population using a predictive model built upon our results should now be carried out to demonstrate the ability of the parameters to discriminate responders from non-responders on a patient by patient basis.

In our study, TNM stage and AJCC classification were not good predictors of therapy response. This could be explained by our suboptimal staging procedure. Since we considered only patients referred for exclusive radiochemotherapy, no patient underwent surgery, and therefore no pathological data were available. Staging was routinely performed using EUS and CT which are known to have limited staging performances [10].

Our present study has limitations. Firstly, we considered a group of only 50 patients with predominantly squamous cell carcinomas since it is the most common histological type of oesophageal cancer in European countries. An analysis based on the tumour histology type considering our patient population did not reveal statistically significant differences, although due to the small number of patients with adenocarcinomas, these results would obviously need to be confirmed. Secondly, our study was inherently limited by its retrospective design and as such some selection bias might be present. However, the treatment regime was homogeneous throughout the recruited patients since all were treated in a single institution. In addition, within this patient population no particular selection criteria were applied. Thirdly, the impact of partial volume effects in the measured SUVs was not assessed in this study. The lack of partial volume correction might have played a role in the reduced predictive value of some of the SUV measurements, although it is unlikely because of the large TVs considered in this work ($40\pm35$ cm$^3$). Lastly, we did consider only primary tumours since the measurements used are simpler to perform in routine clinical practice compared to measurement of overall tumour burden including primary and metastatic lesions. However, given the respective size of metastatic lesions and primary tumours, adding metastatic lesions to the overall TLG would not significantly alter the resulting values and associated conclusions.

## Conclusion

Our results demonstrated that $^{18}$F-FDG baseline image-derived parameters related to the metabolic tumour spatial extent (TL, TV and TLG) are good predictors of response to therapy in oesophageal cancer with sensitivity above 75% and specificity above 85%. Commonly used SUV measurements (max, peak, mean) on the pretreatment $^{18}$F-FDG PET image did not allow statistically significant differentiation of the different response patient groups.

**Acknowledgments** This work was partly funded by ANR (French National Research Agency) under the contract ANR-08-ETEC-005-01.

**Conflicts of interest** None.

## References


1. Parkin DM, Bray F, Ferlay J, Pisani P. Global cancer statistics, 2002. CA Cancer J Clin 2005;55(2):74–108.
2. Hayat MJ, Howlader N, Reichman ME, Edwards BK. Cancer statistics, trends, and multiple primary cancer analyses from the Surveillance, Epidemiology, and End Results (SEER) Program. Oncologist 2007;12(1):20–37.
3. Gebski V, Burmeister B, Smithers BM, Foo K, Zalcberg J, Simes J, et al. Survival benefits from neoadjuvant chemoradiotherapy or chemotherapy in oesophageal carcinoma: a meta-analysis. Lancet Oncol 2007;8(3):226–34.
4. Kelsen DP, Winter KA, Gunderson LL, Mortimer J, Estes NC, Haller DG, et al. Long-term results of RTOG trial 8911 (USA Intergroup 113): a random assignment trial comparison of chemotherapy followed by surgery compared with surgery alone for esophageal cancer. J Clin Oncol 2007;25(24):3719–25.
5. Chirieac LR, Swisher SG, Ajani JA, Komaki RR, Correa AM, Morris JS, et al. Posttherapy pathologic stage predicts survival in patients with esophageal carcinoma receiving preoperative chemoradiation. Cancer 2005;103(7):1347–55.
6. Stahl M, Wilke H, Stuschke M, Walz MK, Fink U, Molls M, et al. Clinical response to induction chemotherapy predicts local control and long-term survival in multimodal treatment of patients with locally advanced esophageal cancer. J Cancer Res Clin Oncol 2005;131(1):67–72.
7. Dragovich T, Campen C. Anti-EGFR-targeted therapy for esophageal and gastric cancers: an evolving concept. J Oncol 2009;2009:804108.
8. Makino T, Yamasaki M, Miyata H, Yoshioka S, Takiguchi S, Fujiwara Y, et al. p53 Mutation status predicts pathological response to chemoradiotherapy in locally advanced esophageal cancer. Ann Surg Oncol 2010;17(3):804–11.
9. Lee JM, Yang SY, Yang PW, Shun CT, Wu MT, Hsu CH, et al. Polymorphism in epidermal growth factor receptor intron 1 predicts prognosis of patients with esophageal cancer after chemoradiation and surgery. Ann Surg Oncol 2011. [Epub ahead of print].
10. van Westreenen HL, Westerterp M, Bossuyt PM, Pruim J, Sloof GW, van Lanschot JJ, et al. Systematic review of the staging performance of 18F-fluorodeoxyglucose positron emission tomography in esophageal cancer. J Clin Oncol 2004;22 (18):3805–12.
11. Krause BJ, Herrmann K, Wieder H, zum Büschenfelde CM. 18F-FDG PET and 18F-FDG PET/CT for assessing response to therapy in esophageal cancer. J Nucl Med 2009;50 Suppl 1:89S–96S.
12. Kwee RM. Prediction of tumor response to neoadjuvant therapy in patients with esophageal cancer with use of 18F FDG PET: a systematic review. Radiology 2010;254(3):707–17.
13. Ott K, Weber WA, Lordick F, Becker K, Busch R, Herrmann K, et al. Metabolic imaging predicts response, survival, and recurrence in adenocarcinomas of the esophagogastric junction. J Clin Oncol 2006;24(29):4692–8.
14. Therasse P, Arbuck SG, Eisenhauer EA, Wanders J, Kaplan RS, Rubinstein L, et al. New guidelines to evaluate the response to treatment in solid tumors. European Organization for Research and Treatment of Cancer, National Cancer Institute of the United States, National Cancer Institute of Canada. J Natl Cancer Inst 2000;92(3):205–16.
15. Larson SM, Erdi Y, Akhurst T, Mazumdar M, Macapinlac HA, Finn RD, et al. Tumor treatment response based on visual and quantitative changes in global tumor glycolysis using PET-FDG imaging. The visual response score and the change in total lesion glycolysis. Clin Positron Imaging 1999;2(3):159–71.
16. Hatt M, Cheze le Rest C, Descourt P, Dekker A, De Ruysscher D, Oellers M, et al. Accurate automatic delineation of heterogeneous functional volumes in positron emission tomography for oncology applications. Int J Radiat Oncol Biol Phys 2010;77(1):301–8.
17. Hatt M, Cheze le Rest C, Turzo A, Roux C, Visvikis D. A fuzzy locally adaptive Bayesian segmentation approach for volume determination in PET. IEEE Trans Med Imaging 2009;28(6):881–93.
18. Schaefer A, Kremp S, Hellwig D, Rübe C, Kirsch CM, Nestle U. A contrast-oriented algorithm for FDG-PET-based delineation of tumour volumes for the radiotherapy of lung cancer: derivation from phantom measurements and validation in patient data. Eur J Nucl Med Mol Imaging 2008;35(11):1989–99.
19. Kruskal W, Wallis W. Use of ranks in one-criterion variance analysis. J Am Stat Assoc 1952;47(260):583–621.
20. Metz CE. Basic principles of ROC analysis. Semin Nucl Med 1978;8(4):283–98.
21. Wahl RL, Jacene H, Kasamon Y, Lodge MA. From RECIST to PERCIST: evolving considerations for PET response criteria in solid tumors. J Nucl Med 2009;50 Suppl 1:122S–50S.
22. Hofman MS, Hicks RJ. Restaging: should we percist without pattern recognition? J Nucl Med 2010;51(12):1830–2.
23. Hatt M, Visvikis D, Albarghach NM, Tixier F, Pradier O, Cheze-le Rest. Prognostic value of (18)F-FDG PET image-based parameters in oesophageal cancer and impact of tumour delineation methodology. Eur J Nucl Med Mol Imaging 2011. [Epub ahead of print].
24. Lucignani G, Larson SM. Doctor, what does my future hold? The prognostic value of FDG-PET in solid tumours. Eur J Nucl Med Mol Imaging 2010;37(5):1032–8.
25. Levine EA, Farmer MR, Clark P, Mishra G, Ho C, Geisinger KR, et al. Predictive value of 18-fluoro-deoxy-glucose-positron emission tomography (18F-FDG-PET) in the identification of responders to chemoradiation therapy for the treatment of locally advanced esophageal cancer. Ann Surg 2006;243(4):472–8.
26. Rizk NP, Tang L, Adusumilli PS, Bains MS, Akhurst TJ, Ilson D, et al. Predictive value of initial PET-SUVmax in patients with locally advanced esophageal and gastroesophageal junction adenocarcinoma. J Thorac Oncol 2009;4(7):875–9.
27. Makino T, Miyata H, Yamasaki M, Fujiwara Y, Takiguchi S, Nakajima K, et al. Utility of response evaluation to neo-adjuvant chemotherapy by (18)F-fluorodeoxyglucose-positron emission tomography in locally advanced esophageal squamous cell carcinoma. Surgery 2010;148(5):908–18.
28. Kato H, Fukuchi M, Miyazaki T, Nakajima M, Tanaka N, Inose T, et al. Prediction of response to definitive chemoradiotherapy in esophageal cancer using positron emission tomography. Anticancer Res 2007;27(4C):2627–33.
29. Yendamuri S, Swisher SG, Correa AM, Hofstetter W, Alani JA, Francis A, et al. Esophageal tumor length is independently associated with long-term survival. Cancer 2009;115(3):508–16.
30. Sillah K, Williams LR, Laasch HU, Saleem A, Watkins G, Pritchard SA, et al. Computed tomography overestimation of esophageal tumor length: implications for radiotherapy planning. World J Gastrointest Oncol 2010;2(4):197–204.
31. Zhong X, Yu J, Zhang B, Mu D, Zhang W, Li D, et al. Using 18F-fluorodeoxyglucose positron emission tomography to estimate the length of gross tumor in patients with squamous cell carcinoma of the esophagus. Int J Radiat Oncol Biol Phys 2009;73(1):136–41.
32. Mamede M, Abreu-E-Lima P, Oliva MR, Nosé V, Mamon H, Gerbaudo VH. FDG-PET/CT tumor segmentation-derived indices of metabolic activity to assess response to neoadjuvant therapy

and progression-free survival in esophageal cancer: correlation with histopathology results. Am J Clin Oncol 2007;30(4):377–88.
33. Roedl JB, Harisinghani MG, Colen RR, Fischman AJ, Blake MA, Mathisen DJ, et al. Assessment of treatment response and recurrence in esophageal carcinoma based on tumor length and standardized uptake value on positron emission tomography-computed tomography. Ann Thorac Surg 2008;86(4):1131–8.
34. Hong TS, Killoran JH, Marmede M, Mamon HJ. Impact of manual and automated interpretation of fused PET/CT data on esophageal target definitions in radiation planning. Int J Radiat Oncol Biol Phys 2008;72(5):1612–8.
35. Arslan N, Miller TR, Dehdashti F, Battafarano RJ, Siegel BA. Evaluation of response to neoadjuvant therapy by quantitative 2-deoxy-2-[18F]fluoro-D-glucose with positron emission tomography in patients with esophageal cancer. Mol Imaging Biol 2002;4(4):301–10.
36. Roedl JB, Colen RR, Holalkere NS, Fischman AJ, Choi NC, Blake MA. Adenocarcinomas of the esophagus: response to chemoradiotherapy is associated with decrease of metabolic tumor volume as measured on PET-CT. Comparison to histopathologic and clinical response evaluation. Radiother Oncol 2008;89(3):278–86.
37. Lee HY, Hyun SH, Lee KS, Kim BT, Kim J, Shim YM, et al. Volume-based parameter of (18)F-FDG PET/CT in malignant pleural mesothelioma: prediction of therapeutic response and prognostic implications. Ann Surg Oncol 2010;17(10):2787–94.
38. Cazaentre T, Morschhauser F, Vermandel M, Betrouni N, Prangère T, Steinling M, et al. Pre-therapy 18F-FDG PET quantitative parameters help in predicting the response to radioimmunotherapy in non-Hodgkin lymphoma. Eur J Nucl Med Mol Imaging 2010;37(3):494–504.
39. Hatt M, Cheze Le Rest C, Albarghach N, Pradier O, Visvikis D. PET functional volume delineation: a robustness and repeatability study. Eur J Nucl Med Mol Imaging 2011;38(4):663–72.
40. Hatt M, Cheze-Le Rest C, Aboagye EO, Kenny LM, Rosso L, Turkheimer FE, et al. Reproducibility of 18F-FDG and 3′-deoxy-3′-18F-fluorothymidine PET tumor volume measurements. J Nucl Med 2010;51(9):1368–76.
41. Hyun SH, Choi JY, Shim YM, Kim K, Lee SJ, Cho YS, et al. Prognostic value of metabolic tumor volume measured by 18F-fluorodeoxyglucose positron emission tomography in patients with esophageal carcinoma. Ann Surg Oncol 2010;17(1):115–22.

Journal of Nuclear Medicine, published on October 11, 2011 as doi:10.2967/jnumed.111.092767

# Impact of Tumor Size and Tracer Uptake Heterogeneity in $^{18}$F-FDG PET and CT Non–Small Cell Lung Cancer Tumor Delineation

Mathieu Hatt[1], Catherine Cheze-le Rest[1], Angela van Baardwijk[2], Philippe Lambin[2], Olivier Pradier[1,3], and Dimitris Visvikis[1]

[1]*INSERM, U650 LaTIM, CHRU Morvan, Brest, France;* [2]*MAASTricht Radiation Oncology Clinic, Maastricht, The Netherlands; and* [3]*Department of Radiotherapy, CHRU Morvan, Brest, France*

The objectives of this study were to investigate the relationship between CT- and $^{18}$F-FDG PET–based tumor volumes in non–small cell lung cancer (NSCLC) and the impact of tumor size and uptake heterogeneity on various approaches to delineating uptake on PET images. **Methods:** Twenty-five NSCLC cancer patients with $^{18}$F-FDG PET/CT were considered. Seventeen underwent surgical resection of their tumor, and the maximum diameter was measured. Two observers manually delineated the tumors on the CT images and the tumor uptake on the corresponding PET images, using a fixed threshold at 50% of the maximum ($T_{50}$), an adaptive threshold methodology, and the fuzzy locally adaptive Bayesian (FLAB) algorithm. Maximum diameters of the delineated volumes were compared with the histopathology reference when available. The volumes of the tumors were compared, and correlations between the anatomic volume and PET uptake heterogeneity and the differences between delineations were investigated. **Results:** All maximum diameters measured on PET and CT images significantly correlated with the histopathology reference ($r > 0.89$, $P < 0.0001$). Significant differences were observed among the approaches: CT delineation resulted in large overestimation (+32% ± 37%), whereas all delineations on PET images resulted in underestimation (from −15% ± 17% for $T_{50}$ to −4% ± 8% for FLAB) except manual delineation (+8% ± 17%). Overall, CT volumes were significantly larger than PET volumes (55 ± 74 cm$^3$ for CT vs. from 18 ± 25 to 47 ± 76 cm$^3$ for PET). A significant correlation was found between anatomic tumor size and heterogeneity (larger lesions were more heterogeneous). Finally, the more heterogeneous the tumor uptake, the larger was the underestimation of PET volumes by threshold-based techniques. **Conclusion:** Volumes based on CT images were larger than those based on PET images. Tumor size and tracer uptake heterogeneity have an impact on threshold-based methods, which should not be used for the delineation of cases of large heterogeneous NSCLC, as these methods tend to largely underestimate the spatial extent of the functional tumor in such cases. For an accurate delineation of PET volumes in NSCLC, advanced image segmentation algorithms able to deal with tracer uptake heterogeneity should be preferred.

Received May 30, 2011; revision accepted Aug. 17, 2011.

For correspondence or reprints contact: Mathieu Hatt, INSERM, U650 LaTIM, CHRU Morvan, 5 Avenue Foch, 29609 Brest, France.

Published online ■■■■.



**Key Words:** NSCLC; $^{18}$F-FDG; tumor delineation; tumor volumes; tumor size; uptake heterogeneity

**J Nucl Med 2011; 52:1–8**
DOI: 10.2967/jnumed.111.092767

The use of $^{18}$F-FDG PET, with the addition of CT since the development of PET/CT devices, has been increasing for staging non–small cell lung cancer (NSCLC) (*1*). In addition, the use of $^{18}$F-FDG PET/CT in radiotherapy treatment planning for the definition of gross tumor volume has been similarly growing (*2*). Manual contouring of the tumor boundaries on the CT images is still the conventional methodology for target volume definition. On the other hand, and despite a high spatial resolution, the delineation on CT alone may be biased by insufficient contrast between tumor and healthy tissues (e.g., in cases of atelectasis, pleural effusion, and fibrosis or for tumors attached to the chest wall or mediastinum). Several studies have investigated the impact of delineation performed on fused $^{18}$F-FDG PET/CT images and have found significant modifications of the treatment plan (size, location, or shape of the gross tumor volume) (*3*) and reduced inter- and intraobserver variability (*4*). Additional benefits from the use of PET relative to CT include the potential to image cellular proliferation and tumor hypoxia using tracers such as 3′-deoxy-3′-$^{18}$F-fluorothymidine and $^{18}$F-fluoromisonidazole or $^{64}$Cu-diacetyl-bis($N^4$-methythiosemicarbazone), respectively.

However, the integration of PET within radiotherapy planning is complex, especially because there is neither consensus nor guidelines regarding the delineation of $^{18}$F-FDG PET tumor uptake or how to subsequently use the delineated functional volumes. Most previously published studies have investigated the use of a specific threshold of PET uptake to define the metabolically active tumor volume (MATV, the tumor volume that can be seen and delineated on an $^{18}$F-FDG PET image) or spatial extent, with a large variability in the recommended threshold and resulting

volumes (*5–8*). A few recent studies have investigated the correlation between tumor histopathology measurements and the threshold of PET uptake (*4,9–12*). For example, the study of Yu et al. (*12*) on 15 patients proposed an optimal threshold of 31% ± 11% of the PET maximum uptake within the tumor for a good correlation with the corresponding histopathology-derived tumor maximum diameter. Considering 3-dimensional reconstructed histopathology volumes instead of only the maximum diameter, Stroom et al. (*10*) recommended a fixed threshold of 42% of the maximum PET uptake based on their findings in a group of 5 patients with rather small tumors. Finally, in the study by Wu et al. (*11*) on 31 patients, 50% of the maximum ($T_{50}$) was proposed as the best threshold for PET uptake delineation in NSCLC with respect to the histopathologic maximum diameter. This conclusion was reached by comparing the results obtained using a range of different fixed thresholds (from 20% to 55%), although only nonstatistically significant differences were found with the other tested values. The same authors subsequently showed that such a threshold was less appropriate than manual delineation, which led to incorrect delineation in some cases (*13*). Manual contouring is far from ideal, as it suffers from large intra- and interobserver variability (*14*) and is also a tedious and time-consuming procedure, especially in 3 dimensions.

Alternatively, other authors have considered the use of adaptive thresholding approaches taking into account the tumor-to-background ratio instead of a fixed threshold but requiring the determination of a background region of interest, as well as optimization for a given scanner model, acquisition protocol, and image reconstruction using phantom acquisitions (*8,15,16*). Using such an approach, van Baardwijk et al. (*4*) obtained a significant correlation with histopathology measurements for 23 NSCLC tumors, as well as reduced interobserver variability. Finally, the use of more advanced image segmentation methodologies to automatically delineate MATV has been proposed in several studies (*17–24*), with variable levels of validation. For example, we have already demonstrated that such automated image segmentation approaches can offer higher accuracy (*18,21*), robustness (*25*), and reproducibility (*14*) than threshold-based (fixed or adaptive) methods.

Some previous studies investigating NSCLC tumor delineation on PET/CT hypothesized a significant influence of the anatomic or metabolic lesion size and activity distribution heterogeneity on both the results and the observed differences between delineation methodologies (*8*). However, those studies neither quantified this heterogeneity nor thoroughly investigated such a correlation with respect to the anatomic tumor and functional uptake sizes. The main objective of our study was therefore to investigate the correlation among anatomic tumor size as determined on CT, the $^{18}$F-FDG uptake level of heterogeneity, and the differences between various automatic PET MATV delineation approaches.

## MATERIALS AND METHODS

### Patient Studies

Twenty-five patients with confirmed NSCLC, stage Ib–IIIb, were included in this study. All patients underwent an $^{18}$F-FDG PET/CT examination for staging purposes before treatment. Patients were instructed to fast for a minimum of 6 h before examination. Free-breathing PET and CT images were acquired 45–60 min after $^{18}$F-FDG injection. A total of seven 5-min bed positions with overlap were used for whole-body PET (Biograph PET/CT; Siemens) acquisitions, which were corrected for attenuation using the CT data and iteratively reconstructed using the ordered-subsets expectation maximization algorithm (4 iterations, 8 subsets). Within a week after PET/CT acquisitions, 17 of the 25 patients underwent surgery (lobectomy), which allowed further macroscopic examination. All specimens were processed in the same way; namely the fresh specimens were put on ice, and 1 pathologist measured the maximum diameter of the tumor in 3 dimensions (*4*). Specimen shrinkage, estimated at about 10%, was not considered since the measurements were performed before fixation in formalin, allowing subsequent immunohistochemical examination, for which the biopsy specimens were paraffin-embedded.

This study was approved by the Institutional Ethics Review Board, and informed written consent was obtained from all patients before their inclusion in the study.

### PET and CT Tumor Delineation

PET images were first up-sampled using a cubic B-spline interpolation scheme (*26*), in such a way that the voxels were of the same size as the associated CT images (Fig. 1). Because the goal of this study was to compare anatomic and MATV as seen and delineated on CT and $^{18}$F-FDG PET images, respectively, manual delineation on fused PET/CT images was not considered. Only primary tumors were delineated on both CT and PET images independently. Tumor anatomic volumes were manually delineated on CT without knowledge of the PET information by 2 observers, both with more than 10 y experience in PET and CT. Functional tumor volumes were manually delineated on PET images by one of the observers (and checked by the other observer) (*13*), as well as using semi- or fully automatic image segmentation tools. A fixed threshold at $T_{50}$ as suggested by Wu et al. (*11*), and an adaptive threshold taking into account the background uptake (*8*), were considered. The adaptive threshold approach was optimized on phantom acquisitions performed on the same PET/CT scanner that was used for the patient acquisitions. The method requires the definition of a manual background region of interest defining the [**Fig. 1**]

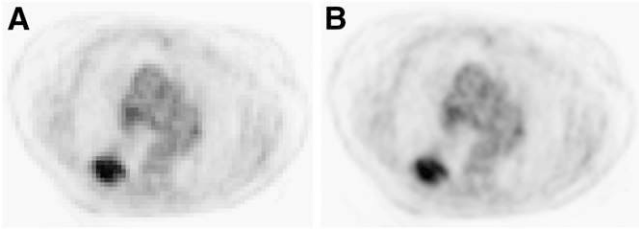



**FIGURE 1.** Illustration of up-sampled PET images (central axial slice). Original PET image with voxel size of 5.31 × 5.31 × 5 mm (A) and PET image up-sampled with voxel size equal to CT (0.98 × 0.98 × 5 mm) (B) using cubic B-spline interpolation.

background uptake to compute a first approximation of the tumor-to-background contrast. Both observers were therefore instructed to place this background region of interest in the lungs, at a distance of several centimeters from the boundaries of the tumors. They were, however, free to choose the actual size and position of the region of interest, which led to 2 different results, denoted A1 and A2. Finally, the fuzzy locally adaptive Bayesian (FLAB) algorithm (*18,21*) was also used. This algorithm allows automatic tumor delineation by computing a probability of belonging to a given class (e.g., tumor or background) for each voxel. This probability is determined by taking into account the voxel intensity with respect to the statistical distributions (characterized by their mean and variance) of the voxels in the various regions of the image, as well as spatial correlation with neighboring voxels. FLAB has demonstrated its ability to accurately differentiate, if necessary, both the overall tumor spatial extent from its surrounding background and the tumor subvolumes with different uptakes (*18*).

### Investigated Parameters and Analysis

First, for the 17 patients for whom macroscopic measurements were available, the maximum diameters were measured as the largest dimension in any orientation, considering the different volume delineations (manual on CT and PET, $T_{50}$, A1 and A2, and FLAB), and compared with the histopathology reference. We reported both absolute (in cm) and relative (%) errors with respect to the maximum diameter to establish a hierarchy between the different methods. Second, for all patients the anatomic tumor volumes defined on CT images and the MATV obtained by each delineation approach were compared with each other. Delineations on original non–up-sampled PET images were performed to verify that the up-sampling would not bias the results of the various methods. Finally, the $^{18}$F-FDG uptake heterogeneity was estimated using the coefficient of variation (COV), defined as the ratio between the SD of the standardized uptake values and the mean standardized uptake value within the delineated MATV. Two different volumes were used to calculate COV. The first was the one obtained using FLAB ($COV_{FLAB}$), since it was found to be the most accurate with respect to histology measurements, whereas the second was the CT-based volume ($COV_{CT}$) copied onto the PET images.

Summary statistics are expressed as mean ± SD. Pearson coefficients were used to estimate correlations between parameters. Paired *t* tests were used to assess the differences between the tumor volume distributions obtained with the various delineation approaches. As most distributions were not normally distributed, they were log-transformed before analysis. All tests were 2-sided, and *P* values of less than 0.05 were considered statistically significant.

## RESULTS

### Comparison with Maximum Diameter (Histopathology Reference)

Table 1 shows the maximum measured diameters of the 17 tumors based on either macroscopic examination or PET and CT images. All measured diameters correlated strongly with macroscopic measurements for all delineation approaches considered ($r$ from 0.89 for $T_{50}$ to 0.99 for FLAB, $P < 0.0001$) (Figs. 2A–2C). Despite high correlations with maximum diameter for all methodologies as shown [**Table 1**] [**Fig. 2**]

**TABLE 1**
Maximum-Diameter Measurements on Pathology and Image Delineations for All 17 Patients

| | Measurement (cm) | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Patient no. | Pathologic | CT1 (manual) | CT2 (manual) | PET (manual) | PET ($T_{50}$) | PET (A1) | PET (A2) | PET (FLAB) |
| 1 | 6.2 | 6.6 | 6.7 | 5.7 | 4.6 | 5 | 4.8 | 5.8 |
| 2 | 2.7 | 3.3 | 3.3 | 3.4 | 2.8 | 3.1 | 2.8 | 3 |
| 3 | 9 | 10.5 | 10.1 | 8.9 | 7 | 7.5 | 7.7 | 9.2 |
| 4 | 1.5 | 1.8 | 1.9 | 2.1 | 1.3 | 1.6 | 1.3 | 1.5 |
| 5 | 1.8 | 3.4 | 3.4 | 2 | 1.2 | 1.4 | 1.3 | 1.6 |
| 6 | 3.1 | 4 | 3.9 | 3.2 | 2.4 | 2.6 | 2.5 | 2.8 |
| 7 | 4.3 | 5 | 5.1 | 4.5 | 3.8 | 3.9 | 3.8 | 3.9 |
| 8 | 3.1 | 5.7 | 5.7 | 5.1 | 2.8 | 4 | 3.7 | 3.5 |
| 9 | 3.5 | 3.9 | 4 | 3.4 | 2.7 | 2.9 | 3 | 3.1 |
| 10 | 5.7 | 7.6 | 7.7 | 7.4 | 7.5 | 4.7 | 6.7 | 5.4 |
| 11 | 5 | 5.1 | 5.3 | 4.7 | 2.7 | 3 | 2.9 | 4.6 |
| 12 | 2.8 | 3.5 | 3.2 | 3.2 | 2.4 | 2.5 | 2.6 | 2.8 |
| 13 | 4.1 | 5.2 | 5.1 | 4.3 | 3.2 | 3.3 | 3.3 | 4 |
| 14 | 4 | 4.8 | 4.9 | 3.7 | 3.2 | 3.4 | 3.2 | 3.9 |
| 15 | 7 | 7.4 | 7.4 | 5.8 | 6.2 | 6.5 | 6.3 | 6.7 |
| 16 | 2.3 | 2.3 | 2.4 | 2.1 | 1.8 | 1.7 | 1.9 | 2.1 |
| 17 | 2.5 | 6 | 5.9 | 4.5 | 2.5 | 2.7 | 2.6 | 2.2 |
| Mean ± SD | 4.0 ± 2.0 | 5.1 ± 2.2 | 5.1 ± 2.1 | 4.2 ± 1.9 | 3.4 ± 1.9 | 3.5 ± 1.6 | 3.6 ± 1.8 | 3.9 ± 2.0 |
| Median | 3.5 | 5.0 | 5.1 | 3.7 | 2.8 | 3.1 | 3.0 | 3.5 |
| Range | 1.5–9 | 1.8–10.5 | 1.9–10.1 | 1.9–8.9 | 1.2–7.5 | 1.4–7.5 | 1.3–7.7 | 1.5–9.2 |
| Pearson *r* | — | 0.90 | 0.91 | 0.95 | 0.89 | 0.95 | 0.93 | 0.99 |
| 95% CI for *r* | — | 0.74–0.96 | 0.76–0.96 | 0.86–0.98 | 0.72–0.96 | 0.85–0.98 | 0.81–0.98 | 0.98–1.00 |

CI = confidence interval.

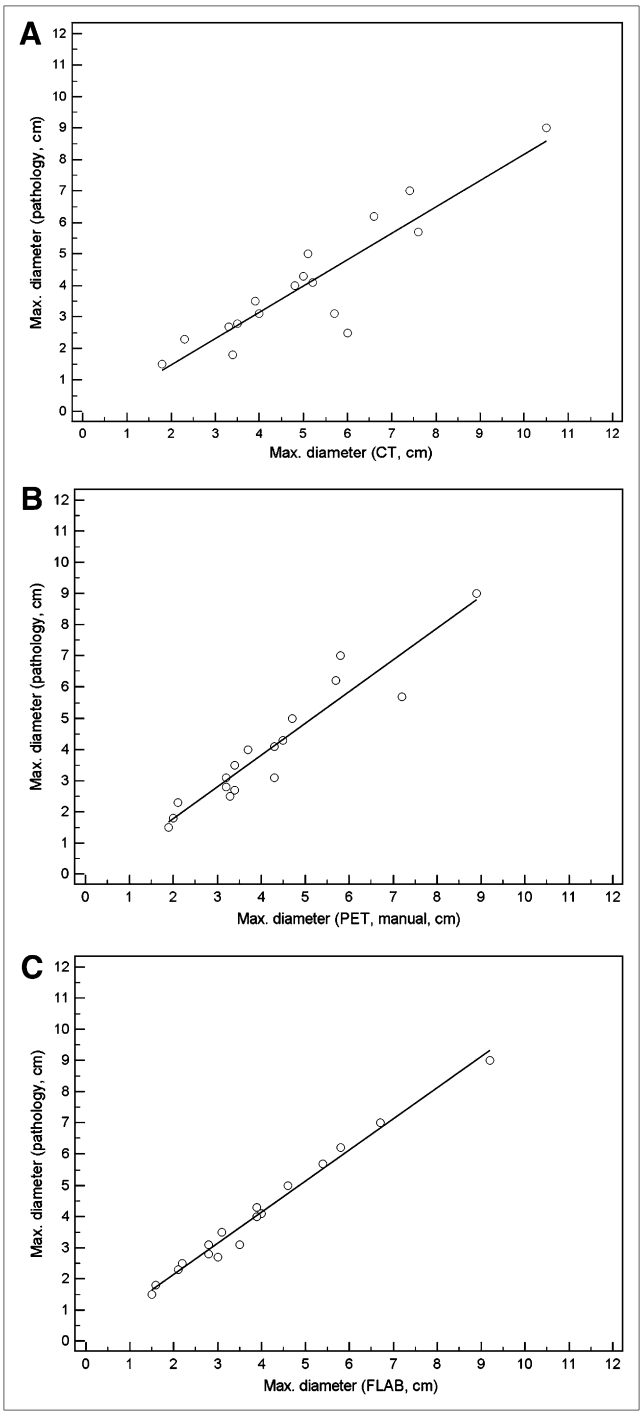



**FIGURE 2.** Correlations with manual delineations on CT (A) and PET (B) and with FLAB delineations on PET (C).

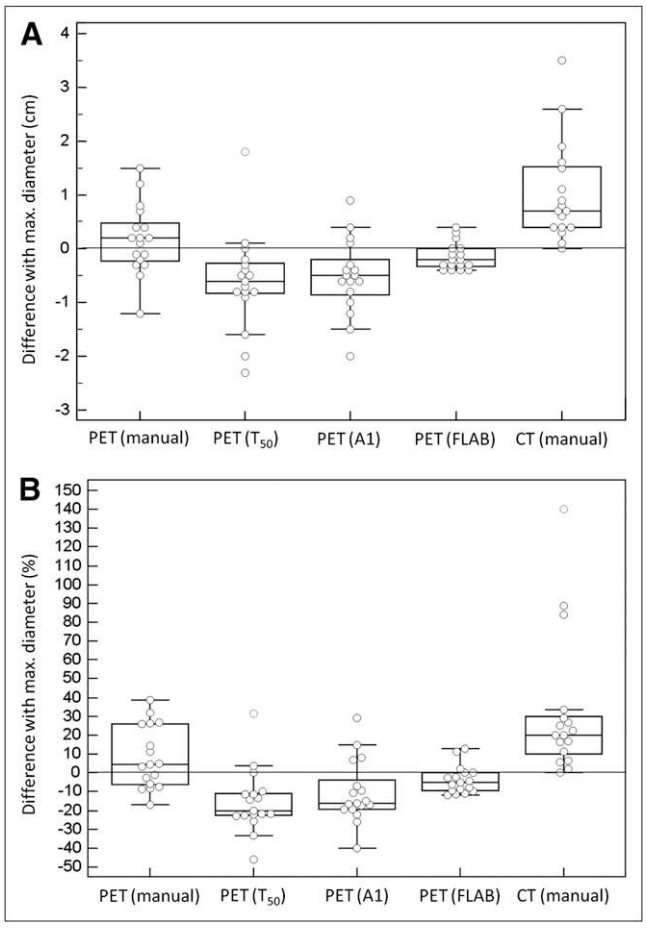



**FIGURE 3.** Absolute (in cm) differences (A) and relative (%) errors (B) between pathology measurements and image-based delineations.

in Table 1 and Figure 2, significant differences were observed among the delineations (Figs. 3A and 3B). On the one hand, CT delineation consistently overestimated the maximum diameter of all tumors (+32% ± 37%), with errors up to 3.5 cm (+140%). Manual delineation on PET images led to no significant bias but a high SD (mean error, 8% ± 17%), with maximum errors of −1.5 cm (−17%) and +1.2 cm (+39%). On the other hand, PET automatic delineations mostly led to underestimation of the real diameter. $T_{50}$ led to the largest underestimation (−15% ± 17%), with errors up to +1.8 cm (+32%) and −2.3 cm (−46%). Adaptive thresholding led to better accuracy, with similar results for both observers (−11% ± 17% and −12% ± 16% for A1 and A2, respectively) and errors up to −2 cm (−40%). FLAB was associated with the most accurate results (−4% ± 8%), with no error above ±0.4 cm (±13%).

[Fig. 3]

### Comparison of Tumor Volumes

Table 2 shows the tumor volumes for all patients. No significant differences in volume determination on CT were found between the 2 observers ($P > 0.08$). Therefore, the results for only 1 observer will be considered. No significant difference was observed between volumes obtained on original or up-sampled PET images.

[Table 2]

Anatomic tumor volumes delineated on CT images were the largest (55 ± 74 $cm^3$) and were significantly different from all volumes defined on PET images ($P < 0.0001$). In addition, all PET-based methodologies resulted in volumes that were significantly different from one another ($P < 0.0001$). Among the PET-defined tumor volumes, and con-

**TABLE 2**
Tumor Volumes Measured on CT and PET Images ($n$ = 25)

| Tumor volume ($cm^3$) ($n$ = 25) | Mean ± SD | Median | Range |
|---|---|---|---|
| CT1 (manual) | 54.5 ± 74.0 | 28.2 | 1.9–338.9 |
| CT2 (manual) | 55.1 ± 74.8 | 29.1 | 1.8–339.4 |
| PET (manual) | 47.3 ± 76.4 | 21.3 | 2.1–356.2 |
| PET ($T_{50}$) | 17.7 ± 25.1 | 9.2 | 8.5–125.8 |
| PET (A1) | 22.6 ± 33.2 | 11.9 | 1.2–166.9 |
| PET (A2) | 21.8 ± 33.9 | 11.3 | 0.9–172.4 |
| PET (FLAB) | 39.5 ± 70.5 | 15.8 | 1.1–345.1 |

sistent with what was observed according to the maximum diameters, the smallest volumes were obtained with $T_{50}$ (18 ± 25 $cm^3$), followed by the adaptive threshold (23 ± 33 $cm^3$), FLAB (40 ± 71 $cm^3$), and manual (47 ± 76 $cm^3$).

Regarding the overlap in delineated volumes, the larger CT volumes almost systematically enclosed the PET-based volumes, except for 8 cases in which small regions of PET uptake were just outside the anatomic volume, involving small margins comprising only a few voxels. The smallest PET uptake volumes generated with $T_{50}$ were also almost systematically enclosed within the volumes defined by the adaptive thresholding, which in turn were mostly enclosed within the FLAB-based volumes. Figure 4 illustrates 3 different cases representative of the various situations encountered.

[**Fig. 4**]

### Correlation of $^{18}$F-FDG Uptake Heterogeneity with Tumor Size and Impact on Delineation

The calculated COVs using the 2 different delineated tumor volumes ($COV_{FLAB}$ and $COV_{CT}$) correlated strongly ($r$ = 0.98, $P < 0.0001$). The heterogeneity of PET uptake in these lung tumors was moderate to high, with a mean $COV_{FLAB}$ of 0.26 ± 0.06 and a range of 0.21–0.48. $COV_{CT}$ was higher, with a mean of 0.37 ± 0.08 (range, 0.3–0.6). Twenty-two of 25 tumors were characterized by a $COV_{FLAB}$ of 0.2–0.3 (0.25–0.4 for $COV_{CT}$), and the 3 most heterogeneous were characterized by a $COV_{FLAB}$ of 0.32, 0.36, and 0.48 (0.46, 0.48, and 0.69, respectively, for $COV_{CT}$). Figure 5 shows 2 different lesions and their associated $COV_{CT}$ and $COV_{FLAB}$. A moderate but significant correlation was found between CT volumes and PET heterogeneity, as larger anatomic volumes exhibited higher heterogeneity ($r$ = 0.44 and $r$ = 0.5 for $COV_{CT}$ and $COV_{FLAB}$, respectively, $P < 0.03$). A similar correlation was found between MATVs and the corresponding heterogeneity, as larger functional volumes also exhibited significantly higher heterogeneity ($r$ = 0.51 and $r$ = 0.58 for $COV_{CT}$ and $COV_{FLAB}$, respectively, $P < 0.002$).

[**Fig. 5**]

Tumor size had an impact on the differences observed between the delineation results using the different images and segmentation approaches considered. A moderate ($r$ = 0.44) correlation was observed between anatomic tumor volumes and the differences between FLAB and $T_{50}$ results (Fig. 6A). The larger the anatomic size of the lesion, the larger were the differences between FLAB and $T_{50}$ volumes ($P$ = 0.025). Similar nonsignificant trends were observed for differences between adaptive thresholding volumes or manual delineation and FLAB ($r < 0.4$, $P > 0.08$). No correlation was found between anatomic tumor size and the differences between CT volumes and all of the PET volumes determined with the different segmentation approaches considered.

[**Fig. 6**]

The impact of PET uptake heterogeneity was more significant than anatomic tumor size on the resulting MATV differences using the PET delineation methodologies considered. As illustrated in Figure 6B, differences between MATV obtained with $T_{50}$ and FLAB correlated strongly ($r < -0.8$) with PET heterogeneity ($P < 0.0001$) estimated either with $COV_{CT}$ or $COV_{FLAB}$. The higher the heterogeneity within the tumor, the smaller was the MATV obtained with $T_{50}$ compared with that derived by FLAB. A similar correlation was observed for the differences between FLAB and A1 ($r < -0.7$, $P < 0.0001$), as well as between FLAB and manual delineation ($r < 0.6$, $P < 0.001$).

## DISCUSSION

Interest in the use of MATV delineation on PET for NSCLC has been growing for several years, especially for radiotherapy applications such as dose redistribution, boosting, and painting, for which MATV is not used in place of anatomic volume but rather as a complement to

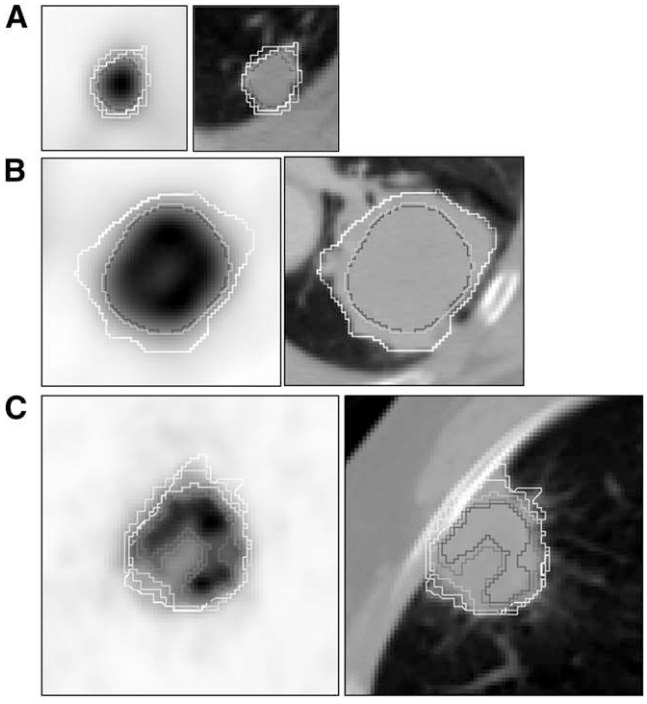



**FIGURE 4.** Small lesions (<2 cm in diameter) (A) and larger lesions with moderate ($COV_{FLAB}$ = 0.23) (B) and higher ($COV_{FLAB}$ = 0.30) (C) heterogeneity. For readability, A1 contours are not shown in B and C and manual PET contours are not shown in B as they were similar to FLAB and $T_{50}$. White = manual on CT; blue = $T_{50}$; purple = A1; green = FLAB.



jnm092767-sn ■ 10/8/11

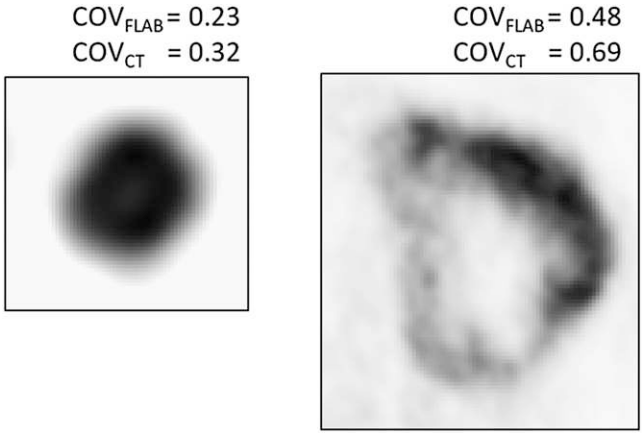



**FIGURE 5.** Heterogeneity estimation for 2 tumors.

increase or redistribute dose within the lesion (*27–29*). These techniques are of interest especially for large tumors characterized by heterogeneous uptake within the MATV. However, the optimal MATV delineation methodology is still subject to debate, especially for these tumor cases.

Our results confirm that large discrepancies can be observed in image-based determination of NSCLC tumor volumes according to the methodology used for tumor delineation. Using morphologic imaging and manual delineation, we saw a large overestimation of tumor volume as previously described by several authors (*13*). Using a fixed threshold of 50% as recommended by Wu et al. (*11*), the estimation of the maximum tumor diameter on PET images was not correct. We observed a constant underestimation of the maximum diameter—a finding that differs from those of Wu et al., who reported mostly overestimations of the maximum diameter of the tumor. This difference is most probably related to the size of the tumors considered in the 2 studies. Wu et al. included mostly small tumors (median diameter, 2 cm; range, 1.1–6.5 cm) whereas we considered larger tumors (4 ± 2 cm; range, 1.5–9 cm). The discordant results could be explained by the failure of binary threshold approaches to deal with heterogeneity, which is more present in larger tumors.

On the other hand, we found differences between CT and PET volumes similar to those found by Wu et al. in their subsequent study (*13*). CT volumes were significantly larger than PET-based volumes in both studies, despite the differences in tumor sizes considered. In our group of patients, we mostly observed that the MATV was completely enclosed in the larger anatomic tumor volumes. Only in a few cases was elevated tracer uptake observed outside the limits of the anatomic tumor, and only a few voxels were involved. This marginal difference may be explained either by imperfect spatial registration between PET and CT or by the impact of respiratory motion.

Using the adaptive thresholding methodology as described by Nestle et al. (*8*), PET tumor sizes did correlate well with the histopathology-based gold standard, albeit with an underestimation of the maximum diameters in our group of lung tumors. Our results agree with those of Van Baardjwick et al. (*4*), who previously investigated a slightly different semiautomatic methodology first proposed by Daisne et al. (*16*).

In the current study, results from the 2 observers using adaptive thresholding were not significantly different, contrary to what was previously observed in the case of esophageal tumors (*30,31*). However, unlike the rather heterogeneous uptake in the mediastinum surrounding esophageal tumors, the lung uptake is more homogeneous, thus leading to negligible variability in the manually determined background values. Manual delineation was less dependent on the heterogeneity within MATV than were threshold-based methods, leading to satisfactory results with no significant bias (mean error $< 10\%$), although there was a large SD (17%) as some MATV were either largely overestimated (mostly the smaller lesions with lower contrast) or underestimated (some of the most heterogeneous ones with complex shapes). Overall, manual delineation correlated strongly with FLAB ($r = 0.96$).

Automatic delineation on PET images using FLAB provided the best estimation of tumor diameters, in accordance with our previous evaluation of FLAB perform-

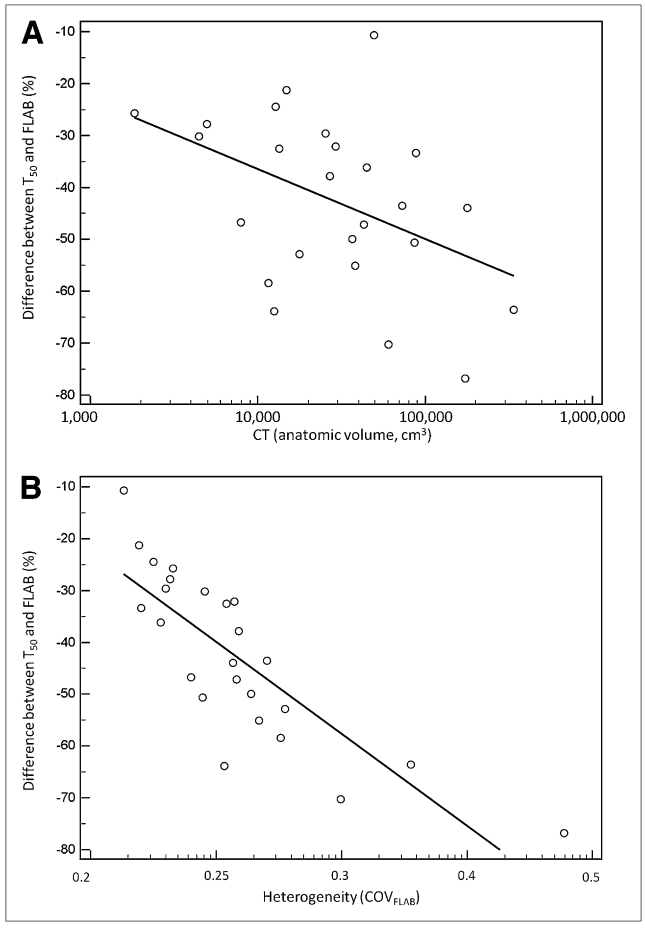



**FIGURE 6.** Correlation between anatomic volume (A) or uptake heterogeneity (B) and differences between $T_{50}$ and FLAB volumes.

ance (*18*). Other advanced segmentation algorithms able to deal with heterogeneous MATV could potentially yield similar satisfactory results (*22,32*). In our previous study, FLAB was compared with a fixed threshold at 42%, instead of 50%, but with similar trends in the observed results. Furthermore, in our previous work the segmentation algorithms were applied to the original PET images without upsampling and therefore with larger voxels. In the present study, resampling was performed for an easier comparison with CT delineations and overlap estimation. This approach resulted in a more accurate estimation of the differences between PET- and CT-based delineation methodologies, without, however, significant differences in the resulting volumes with respect to delineation performed on nonresampled images.

Tracer uptake heterogeneity within the MATV has been recognized as an important factor and a plausible explanation of failed cancer treatments (*33*). Also in malignancies such as sarcomas, esophageal cancer, cervical cancer, and head and neck cancer, studies have shown that local and regional tracer uptake heterogeneity assessment with PET can predict outcome (*34–36*). In NSCLC, Nestle et al. has already observed a larger variability between MATV delineations due to spatial tracer uptake heterogeneity, without, however, quantifying this heterogeneity and the associated correlation with the MATV results (*8*). The impact of heterogeneity on MATV delineation results can be observed and reach statistically significant levels only for objects larger than a few centimeters in diameter, since the limited PET spatial resolution cannot provide accurate imaging of tracer heterogeneity in smaller volumes of interest. These larger tumors are also most frequently encountered in radiotherapy treatment, for which an accurate delineation of the overall MATV may be advantageous, particularly if one considers treatment scenarios involving dose painting or boosting.

Although limited by the small sample of patients and the need to confirm the results in a larger group, our study added several elements to the existing knowledge on the correlation between anatomic tumor size and $^{18}$F-FDG PET uptake in NSCLC. Our results suggest that the larger the tumor, the more heterogeneous the $^{18}$F-FDG PET uptake is likely to be. This suggestion is in agreement with the expected evolution of NSCLC, since necrosis, hemorrhage, or myxoid changes, known to cause areas of low attenuation on CT images, are more likely to appear in larger tumors. A large, heterogeneous MATV is less likely to be accurately delineated using simple fixed or even adaptive binary threshold methods.

In this study, we used the COV to quantify the heterogeneity of PET tracer uptake within the tumor. This heterogeneity factor does not offer any information on the spatial distribution of the heterogeneity and could potentially result in the same value for very different heterogeneous distributions. However, this simple parameter that provides a global measure of heterogeneity is sufficient for the purposes and objectives targeted in this study, allowing us to observe significant correlations between tracer uptake heterogeneity and differences in the MATV segmentation results, either with $COV_{FLAB}$ or $COV_{CT}$. The most heterogeneous lesions were characterized by $COV_{FLAB}$ values above 0.3; however, values from 0.2 to 0.3 were distributed in a rather continuous fashion, making it hard to set a threshold value allowing the differentiation of homogeneous from heterogeneous distributions. A more detailed characterization of the spatial distribution of tumor heterogeneity, which was outside the scope of this study, can be obtained using, for instance, local and regional textural features (*35*).

In studies such as the present one and those published previously within the same context, a common limitation is the lack of respiratory gating. Four-dimensional PET can provide solutions to improve subvolume delineation for dose-painting applications (*37*). However, in our dataset the large size of the tumors should have reduced the potential impact of respiratory motion on the results. In theory, the MATV could have been overestimated for the smallest lesions by both respiratory motion and partial-volume effects. In practice, in our patients only a small fraction of the lesions (10%–20%) were smaller than 2–3 cm.

Finally, a second limitation of our study was the determination of tumor extent based on the measurement of maximum diameter and not the entire volume. Errors in maximum diameter may translate into significantly larger errors with respect to the entire functional volume, especially when heterogeneous uptake distributions are considered. It is indeed possible to obtain an accurate maximum diameter with inaccurate 3-dimensional delineations, especially for complex shapes. Unfortunately full-volume histopathology datasets, for which protocols and corresponding volume estimations are associated with numerous approximations and inaccuracies, are not available yet for NSCLC. Hence, the maximum diameter measurements can be considered as a satisfactory surrogate and have been used in most clinical studies.

## CONCLUSION

Volumes based on CT images were systematically and significantly larger than those based on PET images. In addition, tumor size and PET uptake heterogeneity had a significant impact on the MATV PET delineation results using semi- or fully automatic image segmentation tools. Our results indicate that for a case of large, heterogeneous NSCLC, fixed and adaptive thresholding should not be used for the MATV delineation of $^{18}$F-FDG PET uptake. These methods inherently assume homogeneous uptake in both background and MATV and therefore tend to largely underestimate the spatial extent of the functional tumor in such cases. The use of thresholding approaches should be restricted to smaller lesions with sufficient tumor-to-background contrast or for larger tumors exhibiting homogeneous uptake. For an accurate automatic delineation of MATV in NSCLC, advanced image segmentation algo-

rithms able to deal with tracer uptake heterogeneity should be used.

## DISCLOSURE STATEMENT

The costs of publication of this article were defrayed in part by the payment of page charges. Therefore, and solely to indicate this fact, this article is hereby marked "advertisement" in accordance with 18 USC section 1734.

## ACKNOWLEDGMENTS

This work was partly funded by the French National Research Agency under contract ANR-08-ETEC-005-01. No other potential conflict of interest relevant to this article was reported.

## REFERENCES

1. Hicks RJ, Kalff V, MacManus MP, et al. $^{18}$F-FDG PET provides high-impact and powerful prognostic stratification in staging newly diagnosed non-small cell lung cancer. *J Nucl Med.* 2001;42:1596–1604.
2. MacManus M, Nestle U, Rosenzweig KE, et al. Use of PET and PET/CT for radiation therapy planning: IAEA expert report 2006–2007. *Radiother Oncol.* 2009;91:85–94.
3. Chiti A, Kirienko M, Gregoire V. Clinical use of PET-CT data for radiotherapy planning: what are we looking for? *Radiother Oncol.* 2010;96:277–279.
4. van Baardwijk A, Bosmans G, Boersma L, et al. PET-CT-based auto-contouring in non-small-cell lung cancer correlates with pathology and reduces interobserver variability in the delineation of the primary tumor and involved nodal volumes. *Int J Radiat Oncol Biol Phys.* 2007;68:771–778.
5. Biehl KJ, Kong FM, Dehdashti F, et al. $^{18}$F-FDG PET definition of gross tumor volume for radiotherapy of non-small cell lung cancer: is a single standardized uptake value threshold approach appropriate? *J Nucl Med.* 2006;47:1808–1812.
6. Hellwig D, Graeter TP, Ukena D, et al. $^{18}$F-FDG PET for mediastinal staging of lung cancer: which SUV threshold makes sense? *J Nucl Med.* 2007;48:1761–1766.
7. Yaremko B, Riauka T, Robinson D, Murray B, McEwan A, Roa W. Threshold modification for tumour imaging in non-small-cell lung cancer using positron emission tomography. *Nucl Med Commun.* 2005;26:433–440.
8. Nestle U, Kremp S, Schaefer-Schuler A, et al. Comparison of different methods for delineation of $^{18}$F-FDG PET-positive tissue for target volume definition in radiotherapy of patients with non-small cell lung cancer. *J Nucl Med.* 2005;46:1342–1348.
9. Yu HM, Liu YF, Hou M, Liu J, Li XN, Yu JM. Evaluation of gross tumor size using CT, $^{18}$F-FDG PET, integrated $^{18}$F-FDG PET/CT and pathological analysis in non-small cell lung cancer. *Eur J Radiol.* 2009;72:104–113.
10. Stroom J, Blaauwgeers H, van Baardwijk A, et al. Feasibility of pathology-correlated lung imaging for accurate target definition of lung tumors. *Int J Radiat Oncol Biol Phys.* 2007;69:267–275.
11. Wu K, Ung YC, Hornby J, et al. PET CT thresholds for radiotherapy target definition in non-small-cell lung cancer: how close are we to the pathologic findings? *Int J Radiat Oncol Biol Phys.* 2010;77:699–706.
12. Yu J, Li X, Xing L, et al. Comparison of tumor volumes as determined by pathologic examination and FDG-PET/CT images of non-small-cell lung cancer: a pilot study. *Int J Radiat Oncol Biol Phys.* 2009;75:1468–1474.
13. Wu K, Ung YC, Hwang D, et al. Autocontouring and manual contouring: which is the better method for target delineation using $^{18}$F-FDG PET/CT in non-small cell lung cancer? *J Nucl Med.* 2010;51:1517–1523.
14. Hatt M, Cheze-Le Rest C, Aboagye EO, et al. Reproducibility of $^{18}$F-FDG and 3′-deoxy-3′-$^{18}$F-fluorothymidine PET tumor volume measurements. *J Nucl Med.* 2010;51:1368–1376.
15. Schaefer A, Kremp S, Hellwig D, Rube C, Kirsch CM, Nestle U. A contrast-oriented algorithm for FDG-PET-based delineation of tumour volumes for the radiotherapy of lung cancer: derivation from phantom measurements and validation in patient data. *Eur J Nucl Med Mol Imaging.* 2008;35:1989–1999.
16. Daisne JF, Sibomana M, Bol A, Doumont T, Lonneux M, Gregoire V. Tri-dimensional automatic segmentation of PET volumes based on measured source-to-background ratios: influence of reconstruction algorithms. *Radiother Oncol.* 2003;69:247–250.
17. Yu H, Caldwell C, Mah K, et al. Automated radiation targeting in head-and-neck cancer using region-based texture analysis of PET and CT images. *Int J Radiat Oncol Biol Phys.* 2009;75:618–625.
18. Hatt M, Cheze le Rest C, Descourt P, et al. Accurate automatic delineation of heterogeneous functional volumes in positron emission tomography for oncology applications. *Int J Radiat Oncol Biol Phys.* 2010;77:301–308.
19. El Naqa I, Yang D, Apte A, et al. Concurrent multimodality image segmentation by active contours for radiotherapy treatment planning. *Med Phys.* 2007;34:4738–4749.
20. Montgomery DW, Amira A, Zaidi H. Fully automated segmentation of oncological PET volumes using a combined multiscale and statistical model. *Med Phys.* 2007;34:722–736.
21. Hatt M, Cheze le Rest C, Turzo A, Roux C, Visvikis D. A fuzzy locally adaptive Bayesian segmentation approach for volume determination in PET. *IEEE Trans Med Imaging.* 2009;28:881–893.
22. Belhassen S, Zaidi H. A novel fuzzy C-means algorithm for unsupervised heterogeneous tumor quantification in PET. *Med Phys.* 2010;37:1309–1324.
23. Dewalle-Vignion AS, Betrouni N, Lopes R, Huglo D, Stute S, Vermandel M. A new method for volume segmentation of PET images, based on possibility theory. *IEEE Trans Med Imaging.* 2011;30:409–423.
24. Sebastian TB, Manjeshwar RM, Akhurst TJ, Miller JV. Objective PET lesion segmentation using a spherical mean shift algorithm. *Med Image Comput Comput Assist Interv.* 2006;9:782–789.
25. Hatt M, Cheze Le Rest C, Albarghach N, Pradier O, Visvikis D. PET functional volume delineation: a robustness and repeatability study. *Eur J Nucl Med Mol Imaging.* 2011;38:663–672.
26. Thévenaz P, Blu T, Unser M. Interpolation revisited. *IEEE Trans Med Imaging.* 2000;19:739–758.
27. Thorwarth D, Geets X, Paiusco M. Physical radiotherapy treatment planning based on functional PET/CT data. *Radiother Oncol.* 2010;96:317–324.
28. Petit SF, Aerts HJ, van Loon JG, et al. Metabolic control probability in tumour subvolumes or how to guide tumour dose redistribution in non-small cell lung cancer (NSCLC): an exploratory clinical study. *Radiother Oncol.* 2009;91:393–398.
29. Lambin P, Petit SF, Aerts HJ, et al. The ESTRO Breur Lecture 2009: from population to voxel-based radiotherapy—exploiting intra-tumour and intra-organ heterogeneity for advanced treatment of non-small cell lung cancer. *Radiother Oncol.* 2010;96:145–152.
30. Hatt M, Visvikis D, Pradier O, Cheze-le Rest C. Baseline $^{18}$F-FDG PET image-derived parameters for therapy response prediction in oesophageal cancer. *Eur J Nucl Med Mol Imaging.* 2011;38:1595–1606.
31. Hatt M, Visvikis D, Albarghach NM, Tixier F, Pradier O, Cheze-le Rest C. Prognostic value of $^{18}$F-FDG PET image-based parameters in oesophageal cancer and impact of tumour delineation methodology. *Eur J Nucl Med Mol Imaging.* 2011;38:1191–1202.
32. Nelson AD, Brockway KD, Nelson AS, Piper JW. PET tumor segmentation: validation of a gradient-based method using a NSCLC PET phantom. *J Nucl Med.* 2009;50:340P.
33. Basu S, Kwee TC, Gatenby R, Saboury B, Torigian DA, Alavi A. Evolving role of molecular imaging with PET in detecting and characterizing heterogeneity of cancer tissue at the primary and metastatic sites, a plausible explanation for failed attempts to cure malignant disorders. *Eur J Nucl Med Mol Imaging.* 2011;38:987–991.
34. El Naqa I, Grigsby P, Apte A, et al. Exploring feature-based approaches in PET images for predicting cancer treatment outcomes. *Pattern Recognit.* 2009;42:1162–1171.
35. Tixier F, Le Rest CC, Hatt M, et al. Intratumor heterogeneity characterized by textural features on baseline $^{18}$F-FDG PET images predicts response to concomitant radiochemotherapy in esophageal cancer. *J Nucl Med.* 2011;52:369–378.
36. Eary JF, O'Sullivan F, O'Sullivan J, Conrad EU. Spatial heterogeneity in sarcoma $^{18}$F-FDG uptake as a predictor of patient outcome. *J Nucl Med.* 2008;49:1973–1979.
37. Aristophanous M, Yap JT, Killoran JH, Chen AB, Berbeco RI. Four-dimensional positron emission tomography: implications for dose painting of high-uptake regions. *Int J Radiat Oncol Biol Phys.* 2010;80:900–908.

# Impact of Partial-Volume Effect Correction on the Predictive and Prognostic Value of Baseline $^{18}$F-FDG PET Images in Esophageal Cancer

Mathieu Hatt[1], Adrien Le Pogam[1,2], Dimitris Visvikis[1], Olivier Pradier[1,3], and Catherine Cheze Le Rest[1]

*[1]INSERM, U650 LaTIM, CHRU Morvan, Brest, France; [2]MRC Clinical Sciences Centre, Hammersmith Hospital, London, United Kingdom; and [3]Department of Radiotherapy, CHRU Morvan, Brest, France*

The objective of this study was to investigate the clinical impact of partial-volume effect (PVE) correction on the predictive and prognostic value of metabolically active tumor volume (MATV) measurements on $^{18}$F-FDG PET baseline scans for therapy response and overall survival in esophageal cancer patients. **Methods:** Fifty patients with esophageal cancer treated with concomitant radiochemotherapy between 2004 and 2008 were retrospectively considered. PET baseline scans were corrected for PVE with iterative deconvolution incorporating wavelet denoising. MATV delineation on both original and corrected images was performed using the automatic fuzzy locally adaptive Bayesian methodology. Several parameters were extracted considering the original and corrected images: maximum and peak standardized uptake value (SUV), mean SUV, MATV, and total lesion glycolysis (TLG) (TLG = MATV × mean SUV). The predictive value of each parameter with or without correction was investigated using Kruskal–Wallis tests, and the prognostic value was determined with Kaplan–Meier curves. **Results:** Whereas PVE correction had a significant quantitative impact on the absolute values of the investigated parameters, their clinical value within the clinical context of interest was not significantly modified—a result that was observed for both overall survival and response to therapy. The hierarchy between parameters was the same before and after correction. SUV measurements (maximum, peak, and mean) had nonsignificant ($P > 0.05$) predictive or prognostic value, whereas functional tumor-related measurements (MATV and TLG) were significant ($P < 0.002$) predictors of response and independent prognostic factors. **Conclusion:** PVE correction does not improve the predictive and prognostic value of baseline PET image–derived parameters in esophageal cancer patients.

**Key Words:** esophageal cancer; response to therapy; overall survival; PET; partial volume effects; SUV; tumor volume; total lesion glycolysis

**J Nucl Med 2012; 53:12–20**
DOI: 10.2967/jnumed.111.092775

With a worldwide estimated 5-y survival of only 15% (*1*), esophageal cancer is the third most common malignancy of the digestive tract and is a leading cause of cancer mortality. Its incidence is still increasing, and there is a growing concern regarding its effective management (*2*). Surgical resection remains the most effective treatment; however, many patients have a locally advanced esophageal carcinoma at diagnosis and neoadjuvant therapy before surgery has demonstrated improved survival in such cases (*3*). The maximum improvement in terms of increased overall survival from neoadjuvant treatment is observed for patients who achieve a complete pathologic response (only 15%–30% of cases), with no residual cancer cells in the primary tumor or lymph nodes (*4*). On the other hand, nonresponders (NRs) may be unnecessarily affected by toxicity (*5*). The development of an early diagnostic test offering noninvasive prediction of the response to therapy or survival is therefore of great interest. For tumors that cannot be surgically removed, combined radiochemotherapy is the preferred treatment. In this case too, early assessment of response to therapy would allow a modification in the management of nonresponding patients early during treatment. Such a response assessment becomes even more critical when one considers the availability of new targeted drugs that could be tested with higher efficiency if applied early (*6*).

Along with the standardized uptake values (SUVs) (maximum SUV [$SUV_{max}$] or peak SUV [$SUV_{peak}$]) usually considered in clinical practice, other parameters describing functional lesions—such as metabolically active tumor volume (MATV, defined as the tumor volume that can be seen and delineated on an $^{18}$F-FDG PET image) (*7*), mean SUV ($SUV_{mean}$), and total lesion glycolysis (TLG, defined as the product of MATV and its associated $SUV_{mean}$) (*8*)—have been investigated. The prognostic value of these parameters in esophageal cancer patients for overall or disease-free survival has been demonstrated (*9–12*). On the other hand regarding therapy prediction, several studies on different cancer models have recently suggested using the baseline scan only, instead of the comparison of pretreatment and posttreatment scans (late assessment) or

Received Jul. 6, 2011; revision accepted Sep. 6, 2011.
For correspondence or reprints contact: Mathieu Hatt, LaTIM, INSERM U650, CHRU Morvan, 5 Ave. Foch, 29609 Brest, France.
E-mail: hatt@univ-brest.fr

during-treatment scans (early assessment) (*13*). Such investigations were, for instance, performed in pleural mesothelioma (*14*), non-Hodgkin lymphoma (*15*), and esophageal cancer (*7,16*), demonstrating higher statistical value for MATV-based parameters than SUV measurements, whose predictive value has been found to be conflicting (*17*).

However, in most of these studies, no partial-volume effect (PVE) correction was applied, possibly explaining the observed limited value of SUV. The impact of PVE correction on the clinical value of SUV measurements has been investigated by a limited number of authors. Hoetjes et al. (*18*) investigated the impact of 4 PVE correction strategies on 15 breast cancer patients, regarding the early metabolic PET response after 1 cycle of chemotherapy. The SUV decrease between the pretreatment scan and the scan early during treatment was found to be lower after PVE correction (26%–27% vs. 31%) for the first 3 methods but not for the fourth one based on binary tumor masks (30%). Van Heijl et al. (*19*) recently demonstrated a nonsignificant impact of PVE correction on the correlation between disease-free survival and $^{18}$F-FDG PET SUV measurements in 52 esophageal cancer patients. In this study, a PVE correction method based on binary tumor masks generated with adaptive thresholding delineation was used, and disease-free survival was the only clinical endpoint investigated. Both the use of adaptive thresholding and the PVE correction method based on tumor masks assume a homogeneous tracer distribution in both tumor and background and are therefore likely to provide only approximate correction (*20*). On the other hand, no data are currently available regarding the impact of PVE correction on the value of baseline $^{18}$F-FDG PET–based measurements for the prediction of overall survival and response to therapy in esophageal cancer.

The current study was therefore performed to investigate the impact of an advanced PVE correction methodology and the use of an accurate MATV delineation approach on both the predictive and the prognostic value of baseline $^{18}$F-FDG PET scan–derived parameters.

## MATERIALS AND METHODS

### Patients

Fifty consecutive patients with newly diagnosed esophageal cancer were included and retrospectively analyzed. The characteristics of the patients are given in Table 1. Most of the patients (45 of 50) were men, aged 65 ± 9 y at the time of diagnosis. Seventy-four percent of the tumors originated from the middle and lower esophagus, and 72% were squamous cell carcinoma. None of the patients underwent surgery, and all were treated with concomitant radiochemotherapy between 2004 and 2009. The therapy regime included 3 courses of 5-fluorouracil and cisplatin and a median radiation dose of 60 Gy given in 180-cGy fractions delivered once daily, 5 d a week for 6–7 wk. As part of the routine procedure for the initial staging in esophageal cancer, each patient was referred for an $^{18}$F-FDG PET study before treatment, and these baseline scans were used in this study.

**TABLE 1**
Patient Demographics and Clinical Characteristics

| Parameter | No. of patients ($n$ = 50) |
|---|---|
| Sex | |
| Male | 45 (90) |
| Female | 5 (10) |
| Site | |
| Upper esophagus | 13 (26) |
| Middle esophagus | 20 (40) |
| Lower esophagus | 17 (34) |
| Histology type | |
| Adenocarcinoma | 14 (28) |
| Squamous cell carcinoma | 36 (72) |
| Histologic differentiation | |
| Well differentiated | 14 (28) |
| Moderately differentiated | 12 (24) |
| Poorly differentiated | 5 (10) |
| Unknown | 19 (38) |
| TNM stage | |
| T1 | 7 (14) |
| T2 | 8 (16) |
| T3 | 24 (48) |
| T4 | 11 (22) |
| N0 | 20 (40) |
| N1 | 30 (60) |
| M0 | 34 (68) |
| M1 | 16 (32) |
| American Joint Committee on Cancer stage | |
| I | 4 (8) |
| IIA | 8 (16) |
| IIB | 6 (12) |
| III | 16 (32) |
| IVA | 16 (32) |

Age range of patients was 45–84 y, and median was 69 y. Data in parentheses are percentages.

Overall survival was determined as the time between initial diagnosis and last follow-up or death. Response to therapy was evaluated 1 mo after the completion of the concomitant radiochemotherapy using conventional thoracoabdominal CT and endoscopy. Patients were classified as NRs (including stable and progressive disease), partial responders (PRs), or complete responders (CRs). Response evaluation was based on CT evolution between pretreatment and posttreatment scans using response evaluation criteria in solid tumors (*21*). Patients also underwent fibroscopy in the case of partial or complete response. Complete response was confirmed by the absence of visible disease in the endoscopy and no viable tumor on biopsy. Partial CT response was confirmed by macroscopic residual (disease >10% viable) on biopsy. No discordance was observed between pathologic, when available, and CT evaluation. The current analysis was performed after an approval by the institutional ethics review board.

### $^{18}$F-FDG PET Acquisitions

$^{18}$F-FDG PET studies were performed before the treatment. Patients were instructed to fast for at least 6 h before an injection of $^{18}$F-FDG (5 MBq/kg). Static emission images were acquired from head to thigh beginning 60 min after injection and with 2 min per bed position, on a Gemini PET/CT system (Philips). Images were reconstructed using the row-action

maximum-likelihood 3-dimensional algorithm according to standard clinical protocol: 2 iterations, relaxation parameter of 0.05, 5-mm 3-dimensional gaussian postfiltering, a $4 \times 4 \times 4$ mm voxel grid sampling, and attenuation correction based on a low-dose CT scan.

### PET Image PVE Correction and Image Analysis

Images were corrected for PVE using an iterative deconvolution methodology that has been previously validated (*22*). Its principle is to iteratively estimate the inversion of the scanner's point spread function, which is assumed to be known and spatially invariant in the field of view. The considered lesions were all in the same body region, and this approximation should therefore not significantly affect the applied correction on a patient-by-patient comparison basis. Iterative deconvolution methods, such as those of Lucy-Richardson (L-R) (*23,24*) or Van Cittert (*25*), are known for the amplification of noise associated with an increasing number of iterations. To solve this issue, wavelet-based denoising of the residual was introduced within the iterative L-R deconvolution using Bayeshrink filtering (*26*), leading to images corrected for PVE without significant noise addition. The following are advantages of this methodology: it is able to generate entire whole-body corrected images independently of any manual or automatic segmentation of regions of interest, and it is voxel-based and therefore does not assume homogeneous regional radiotracer distributions for the tumor or surrounding background.

### Tumor Delineation and Parameter Extraction

For each patient, the tumor was identified on the baseline pretreatment PET images by an experienced nuclear physician. It was then delineated using the fuzzy locally adaptive Bayesian algorithm (*20,27*) on both the original (without PVE correction) and the PVE-corrected images. This segmentation approach has been shown to give both robust and reproducible functional volume delineations under variable image noise characteristics (*28,29*).

The following parameters were subsequently extracted from each baseline image with or without correction for PVE: $SUV_{max}$, $SUV_{peak}$ (defined as the mean of $SUV_{max}$ and its 26 neighbors [roughly corresponding to a 1-cm region of interest]), $SUV_{mean}$ within the volume, MATV, and TLG (determined by multiplying $SUV_{mean}$ with the corresponding MATV).

### Statistical Analysis

Pearson coefficients were used to estimate correlation between the image-derived parameters, and paired *t* tests were used to characterize the differences between uncorrected and corrected parameters. The correlation between response to therapy and each parameter was investigated using the Kruskal–Wallis test as a nonparametric statistic allowing the comparison of parameter distributions associated with each category of response (CR, PR, and NR). This test does not assume a normal distribution of variables, and the computation of its statistic H is based on ranks instead of absolute values of variables (*30*). Regarding survival, for each considered parameter, Kaplan–Meier survival curves were generated (*31*) for which the most discriminating threshold value allowing differentiation of the groups of patients was identified using receiver-operating-characteristic methodology (*32*). The prognostic value of each parameter in terms of overall survival was assessed by the log-rank test.

The significance of the following factors (with or without correction) was tested: $SUV_{max}$, $SUV_{peak}$, MATV, $SUV_{mean}$, and TLG. All tests were performed 2-sided using the MedCalc statistical software (MedCalc Software), and *P* values below 0.05 were considered statistically significant.

## RESULTS

### Impact of PVE Correction on Image-Derived Parameters

The PVE correction affected the images that could be assessed visually, with a higher contrast between the tumor and the surrounding tissues, as can be seen in Figure 1 and is illustrated using profiles in Figure 2. Table 2 provides the distributions of volumes and associated parameters measured in original and corrected images.

MATVs delineated on original images and images corrected for PVE were highly correlated ($r > 0.998$; confidence

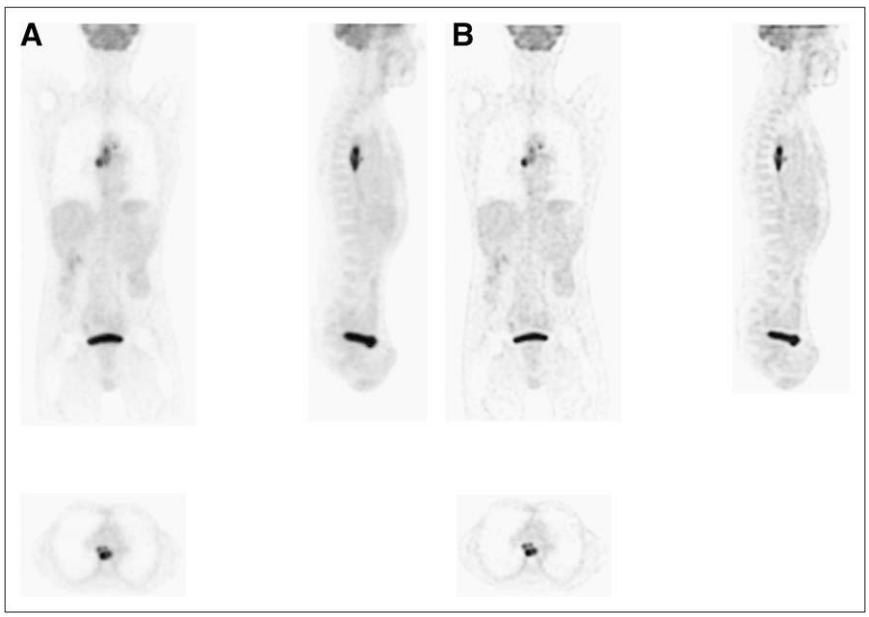



**FIGURE 1.** Illustration of iterative deconvolution PVE correction on whole-body $^{18}$F-FDG PET image, with original image (A) and corrected image (B).

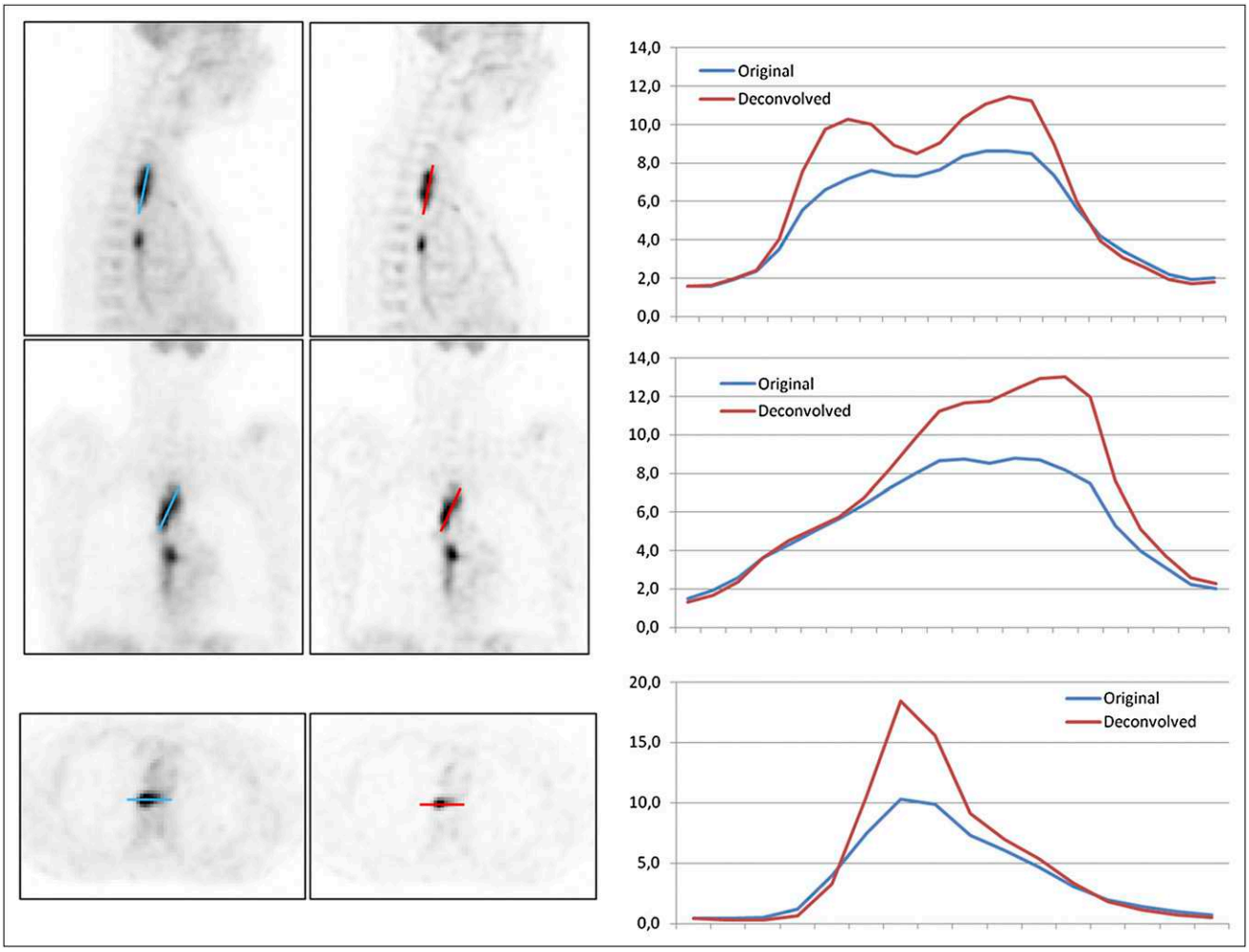



**FIGURE 2.** Qualitative differences between original and corrected PET images of esophageal lesion of MATV above 25 $cm^3$ using profiles on axial, sagittal, and coronal planes.

interval, 0.997–0.999; $P < 0.0001$). However, MATVs delineated on PVE-corrected images were systematically smaller ($P < 0.001$) by on average $-10\% \pm 5\%$ (range, $-1.5\%$ to $-22.4\%$), resulting in a mean volume difference of $-4 \pm 3$ $cm^3$ ($40 \pm 36$ $cm^3$ vs. $36 \pm 34$ $cm^3$). This difference is illustrated on 3 different tumors in Figure 3. There was no significant correlation between these differences and the PET lesion volumes ($r < 0.2$, $P > 0.18$).

All primary lesions were detected by $^{18}$F-FDG PET and exhibited a rather high uptake with a mean $SUV_{max}$ of $10 \pm 4$. As expected, $SUV_{peak}$ and $SUV_{mean}$ measurements were comparatively lower ($8 \pm 3$ and $6 \pm 2$, respectively). All SUV measurements are summarized in Table 2. After iterative deconvolution, $SUV_{max}$, $SUV_{peak}$, and $SUV_{mean}$ were $15 \pm 6$, $10 \pm 4$, and $7 \pm 3$, respectively. All were significantly higher than noncorrected values ($P < 0.05$). $SUV_{max}$ increased by $54\% \pm 23\%$ (range, 18%–157%), whereas the impact on $SUV_{peak}$ and $SUV_{mean}$ was lower, with a mean increase of $27\% \pm 10\%$ (range, 8%–51%) and $28\% \pm 11\%$ (range, 9%–59%), respectively. Considering the PVE correction–induced decrease of MATV ($-10\% \pm 5\%$) and increase of corresponding $SUV_{mean}$

**TABLE 2**
Distributions of Parameters With and Without PVE Correction

| Definition | Notation | Original mean ± SD | PVE correction mean ± SD |
|---|---|---|---|
| Highest SUV | $SUV_{max}$ | 9.7 ± 3.9 | 14.9 ± 6.1 |
| Mean of $SUV_{max}$ and its 26 neighbors | $SUV_{peak}$ | 8.0 ± 3.3 | 10.1 ± 4.0 |
| $SUV_{mean}$ within MATV | $SUV_{mean}$ | 5.8 ± 2.4 | 7.4 ± 3.1 |
| MATV ($cm^3$) | MATV | 39.9 ± 36.1 | 36.2 ± 33.7 |
| Total lesion glycolysis (g) | TLG | 218.1 ± 208.3 | 235.8 ± 218.1 |

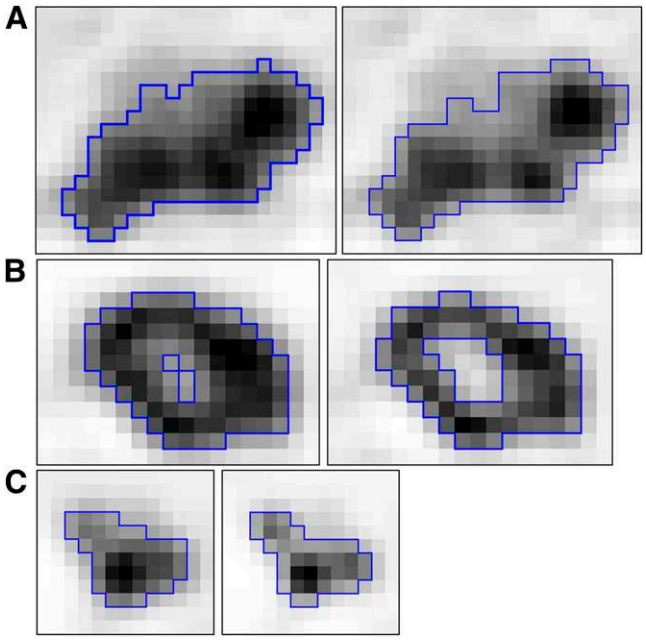



**FIGURE 3.** Examples of fuzzy locally adaptive Bayesian delineation results (blue contours) on original (left) and corrected (right) PET images with large, slightly heterogeneous MATV (A); MATV with necrotic core (B); and small, homogeneous MATV (C).

(+28% ± 11%), PVE correction resulted in significantly higher TLG values (+14% ± 12%; range, −2 to +50%) ($P < 0.0001$).

The increases of $SUV_{max}$ and $SUV_{peak}$ after PVE correction did not correlate with MATV ($r < 0.2$, $P > 0.2$), whereas the increase of $SUV_{mean}$ correlated inversely with MATV ($r = -0.79$, $P < 0.0001$), with higher increases observed for smaller volumes.

### Impact of PVE Correction on Predictive and Prognostic Values

Twenty-five patients were classified as PR, 11 were CR, and 14 were NR (including stable and progressive disease). With a median follow-up of 60 mo (range, 10–84 mo), the median overall survival was 12 mo and the 1-y and 2-y survival rates were 60% and 35%, respectively. At the time of last follow-up, 10 patients were alive with no evidence of disease, 9 were alive with recurrent disease, and 31 had died. Survival was significantly correlated with response, as overall survival was 24 ± 15 (median, 21), 22 ± 20 (median, 14), and 9 ± 4 (median, 10) months for CR, PR, and NR, respectively ($P < 0.01$). Results concerning the prognostic and predictive values of all considered parameters with and without PVE correction are summarized in Tables 3 and 4.

Initial $SUV_{max}$, whether corrected for PVE or not, was not predictive of response to therapy ($P = 0.2$ and 0.3 for $SUV_{max}$ and $SUV_{max}$ with PVE correction, respectively), although CRs tend to have a smaller $SUV_{max}$ (7.8 ± 4.2 and 12.2 ± 6.6 after PVE correction) than PRs and NRs (10.2 ± 3.7 and 10.3 ± 3.8 for PR and NR, respectively, and 15.9 ± 6.0 and 15.5 ± 5.7, respectively, after PVE correction) (Fig. 4A). $SUV_{peak}$ led to slightly more differentiated groups of response without reaching statistical significance ($P = 0.08$), with a mean value of 6.2 ± 3.6 in CRs, whereas both PRs and NRs were characterized by a similarly higher $SUV_{peak}$ (8.5 ± 3.1 and 8.5 ± 3.2 for PRs and NRs, respectively). After PVE correction, the results using $SUV_{peak}$ were similar, with 7.8 ± 4.4, 10.7 ± 3.7, and 10.8 ± 3.9 for CRs, PRs, and NRs, respectively ($P = 0.1$). The $SUV_{mean}$ measurements could not significantly predict response ($P = 0.07$), and the differentiation between the 3 groups of response considered on the basis of $SUV_{mean}$ was still not possible after PVE correction ($P > 0.14$).

None of the SUV measurements was a significant prognostic factor in the univariate analysis, despite a trend for longer survival associated with lower SUV (maximum, peak, or mean). For instance, an $SUV_{max}$ below a threshold of 8 or an $SUV_{mean}$ under 6.5 tend to be associated with a better outcome and a median survival of 20 versus 13 mo

**TABLE 3**
Kruskal–Wallis Test Results

| Parameter | H | *P* | Response differentiation? ($P < 0.05$) | | |
|---|---|---|---|---|---|
| | | | CR ($n = 11$)/NR ($n = 14$) | CR ($n = 11$)/PR ($n = 25$) | PR ($n = 25$)/NR ($n = 14$) |
| $SUV_{max}$ | 3.6 | 0.17 | No | No | No |
| $SUV_{max}$ with PVE correction | 2.4 | 0.31 | No | No | No |
| $SUV_{peak}$ | 5.1 | 0.08 | No | No | No |
| $SUV_{peak}$ with PVE correction | 4.7 | 0.10 | No | No | No |
| $SUV_{mean}$ | 5.5 | 0.07 | No | No | No |
| $SUV_{mean}$ with PVE correction | 3.9 | 0.14 | No | No | No |
| MATV ($cm^3$) | 20.7 | <0.0001 | Yes | Yes | Yes |
| MATV with PVE correction ($cm^3$) | 20.7 | <0.0001 | Yes | Yes | Yes |
| TLG (g) | 25.1 | <0.0001 | Yes | Yes | Yes |
| TLG with PVE correction (g) | 25.2 | <0.0001 | Yes | Yes | Yes |

H statistic and associated *P* value are given for each parameter, with ability to differentiate ($P < 0.05$) each pair of response groups among patients.

**TABLE 4**
Univariate Analysis Results Using Kaplan–Meier Survival Curves

| Parameter | Threshold | HR | HR 95% confidence interval | *P* | Median survival (mo) |
|---|---|---|---|---|---|
| $SUV_{max}$ | 8 | 1.5 | 0.7–3.1 | 0.28 | 20 vs. 13 |
| $SUV_{max}$ with PVE correction | 11 | 1.6 | 0.7–3.2 | 0.26 | 20 vs. 13 |
| $SUV_{peak}$ | 7 | 1.4 | 0.7–2.8 | 0.31 | 16 vs. 10 |
| $SUV_{peak}$ with PVE correction | 9 | 1.8 | 0.9–3.6 | 0.11 | 20 vs. 11 |
| $SUV_{mean}$ | 6.5 | 1.7 | 0.8–3.6 | 0.15 | 16 vs. 10 |
| $SUV_{mean}$ with PVE correction | 7.5 | 1.7 | 0.8–3.5 | 0.12 | 20 vs. 10 |
| MATV ($cm^3$) | 85 | 3.9 | 1.0–15.2 | 0.0004 | 20 vs. 6 |
| MATV with PVE correction ($cm^3$) | 80 | 3.4 | 0.9–11.7 | 0.0024 | 16 vs. 10 |
| TLG (g) | 260 | 2.9 | 1.2–6.8 | 0.0012 | 21 vs. 10 |
| TLG with PVE correction (g) | 280 | 3.2 | 1.3–7.6 | 0.0004 | 21 vs. 10 |

($P = 0.3$) and 16 versus 10 mo ($P = 0.15$), respectively. Similarly, after PVE correction no threshold value could significantly differentiate groups of patients regarding their survival (Figs. 5A and 5B).

Contrary to SUV measurements with or without PVE correction, the parameters related to functional volume (MATV and TLG) allowed significant ($P < 0.0001$) differentiation of the 3 response groups and were significant prognostic factors ($P < 0.002$), as illustrated in Figure 4C. No significant differences were found using the original or PVE-corrected values.

The parameter that allowed for the best differentiation of patient groups was the TLG ($P < 0.0001$). CRs were characterized by a TLG of 55 ± 45 g, whereas PRs and NRs had a TLG of 178 ± 143 and 416 ± 238 g, respectively. After PVE correction, the absolute values of each group rose to 62 ± 45, 200 ± 155, and 437 ± 249 g for CRs, PRs, and NRs, respectively, leading to the same discrimination between groups of response ($P < 0.0001$). Although slightly less efficient than TLG, the use of MATV allowed a statistically significant differentiation of the 3 response groups ($P < 0.0001$). Use of the MATV values extracted from PVE correction images led to exactly the same discriminating power ($P < 0.0001$).

MATV and TLG were also good prognostic factors, with high MATV and TLG values being significantly associated with shorter survival, with hazard ratios between 3 and 4 (Table 3). A MATV above 85 $cm^3$ was identified as a predictor of poor outcome, with a median survival of only 6 mo, versus 20 mo for patients with a smaller MATV ($P = 0.0004$), as illustrated in Figure 5C. In addition, a MATV below 15 $cm^3$ was associated ($P = 0.009$) with longer survival (49 mo) than a larger MATV (11 mo). Similar results were obtained using the MATVs measured on the PVE-corrected images, with a median survival of 20 mo for patients with tumor volume with PVE correction below 80 $cm^3$ versus 10 mo for patients with MATV above 80 $cm^3$ ($P < 0.002$). Regarding TLG, a threshold of 260 g was found to be a good discriminating factor for outcome (21 vs. 10 mo, $P = 0.0012$), whereas using PVE-corrected TLG led to similar results, with a slightly higher threshold (TLG with PVE correction = 280 g, 21 vs. 10 mo, $P = 0.0004$).

## DISCUSSION

Our study investigated the impact of PVE correction on the predictive and prognostic values of different parameters derived using the baseline pretreatment PET images. Our results confirmed that PVE correction significantly affects quantitative SUVs, with an average increase of above 50% for $SUV_{max}$, in agreement with previous studies (*18,19*), and a lower increase (<30%) for $SUV_{peak}$ and $SUV_{mean}$. The lower increase observed for $SUV_{peak}$ and $SUV_{mean}$ is related to the fact that the L-R deconvolution is a voxel-by-voxel process leading to enhancement of contrasts between subvolumes within the MATV and both lower- and higher-voxels SUVs included in the averaging associated with the calculation of $SUV_{mean}$ and $SUV_{peak}$. PVE correction did not significantly affect the delineation of the MATV. Overall, MATVs delineated on the corrected images were only slightly smaller than those determined on the original images. The mean reduction of 10% was within the reproducibility limits of confidence intervals regarding tumor volume measurements on double-baseline PET scans using fuzzy locally adaptive Bayesian algorithm method (±30%) (*29*). This limited impact of PVE correction on MATV can be explained by the fact that PVE is dependent on tumor size and is more pronounced on small lesions (*33*). In our group of patients, the tumors were rather large (40 ± 30 $cm^3$); therefore, the relative variation of volumes with respect to the entire volume is small. Twelve patients (25%) had an MATV of around 10 $cm^3$ or smaller. In addition, the use of a robust delineation approach instead of threshold-based methods in various configurations of blur and noise (*28,34*) ensured a limited variability in the MATV delineation results between original and corrected images.

As previously demonstrated (*7,12*), MATV and TLG extracted from noncorrected $^{18}$F-FDG PET pretreatment acquisitions had high clinical value. In contrast, none of the usual SUV measurements (maximum, peak, or mean) considered in clinical practice was significantly associated with therapy response or survival, as also reported in the 2 largest available prospective trials (*35,36*).

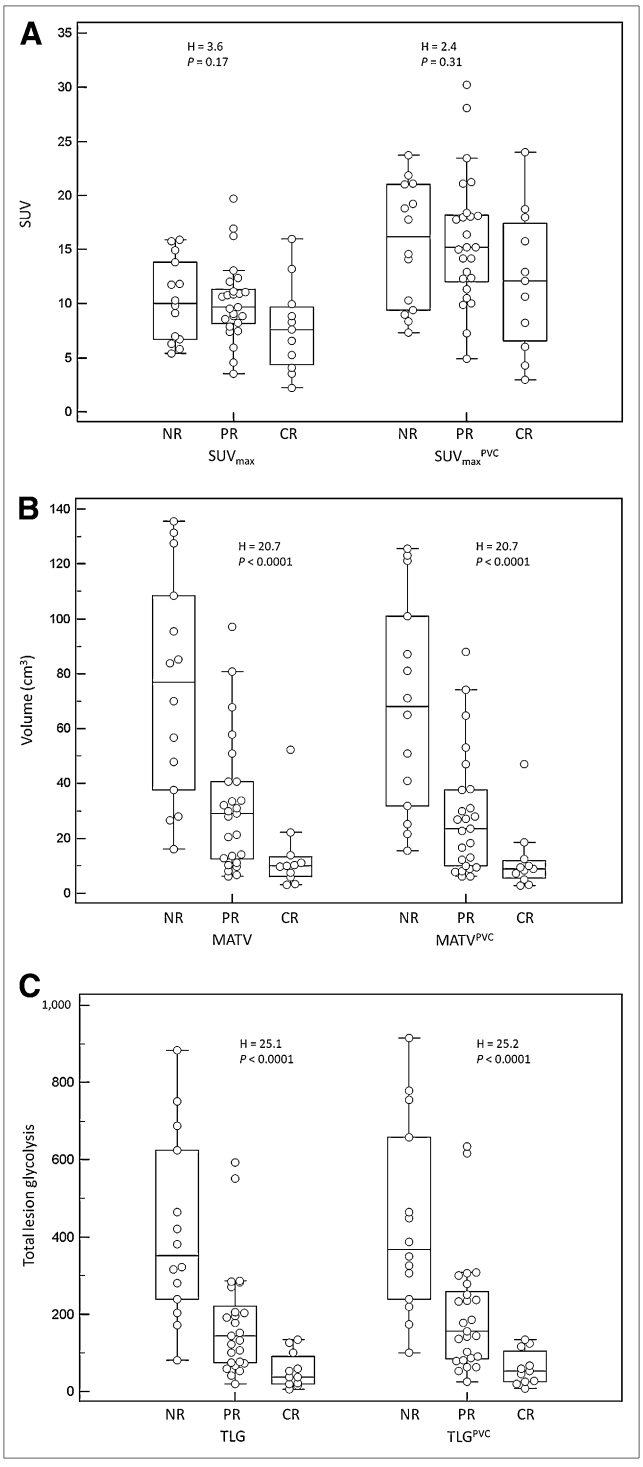



**FIGURE 4.** Examples of distributions of NRs, PRs, and CRs and associated Kruskal–Wallis test results: $SUV_{max}$ and $SUV_{max}$ with PVE correction (A), MATV and MATV with PVE correction (B), and TLG and TLG with PVE correction (C). $MATV^{PVC}$ = MATV with PVE correction; $SUV_{max}^{PVC}$ = SUV with PVE correction; $TLG^{PVC}$ = TLG with PVE correction.

Regarding response to therapy prediction using SUVs, we found that PVE correction did not improve the already demonstrated low discriminating power of any of the SUV measurements considered (*7*). This can be explained by the combination of several factors. First, without PVE correction, the trend of low SUV being associated with better outcome may have been exaggerated by an underestimation of SUV, because CRs had also smaller volumes in addition to low $SUV_{max}$. Second, after PVE correction all 3 response groups had increased $SUV_{max}$ but with still no significant

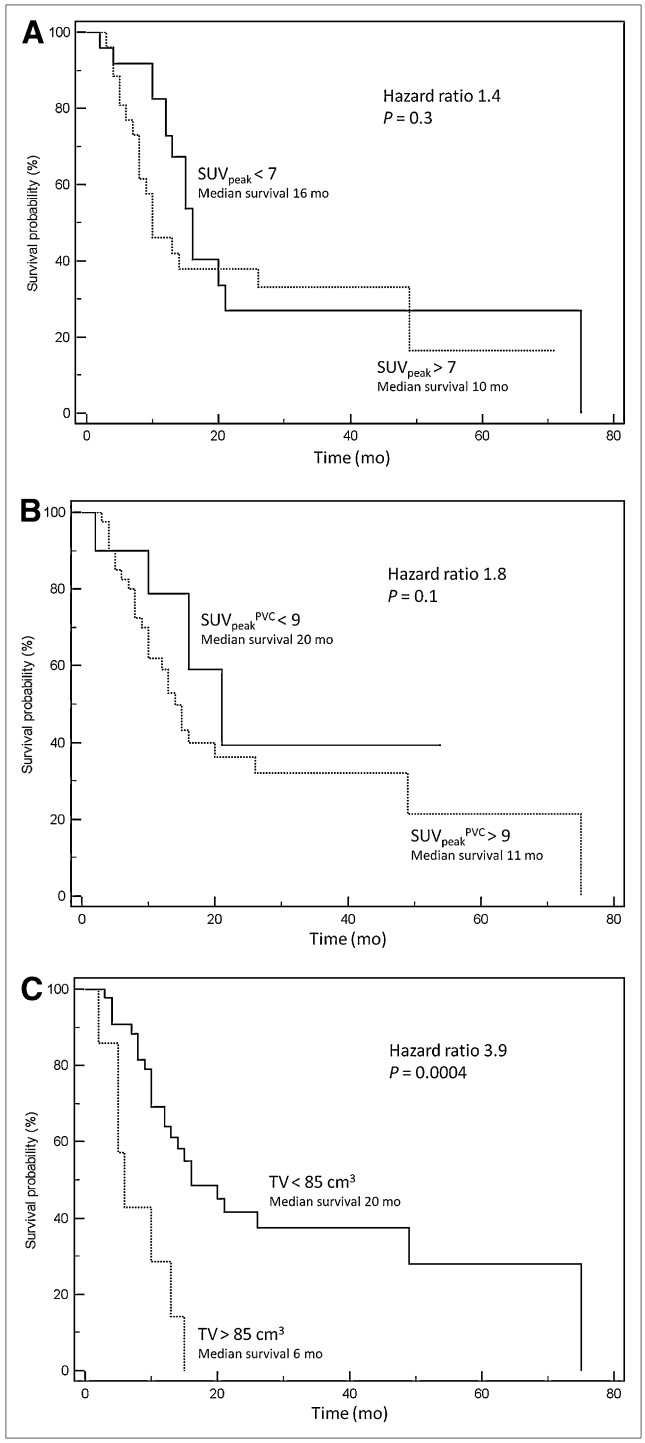



**FIGURE 5.** Examples of Kaplan–Meier survival curves obtained using $SUV_{peak}$ (A), $SUV_{peak}$ with PVE correction (B), and MATV (C). $SUV_{peak}^{PVC}$ = $SUV_{peak}$ with PVE correction.

difference between the groups. We have demonstrated that $SUV_{mean}$ increase after PVE correction was inversely correlated with tumor volume ($r = 0.8$, $P < 0.0001$), with smaller volumes being characterized by higher $SUV_{mean}$ increases after PVE correction than larger volumes. The $SUV_{mean}$ within the MATV of PRs and NRs was therefore increased by a smaller amount (+20% ± 9%) than those within the MATV of CRs (+34% ± 13%), which were associated with smaller tumor volumes. The mean tumor SUVs of CRs were therefore closer to the $SUV_{mean}$ of PRs and NRs after correction. Hence, the discriminating power of $SUV_{mean}$ was reduced by PVE correction. A similar trend was observed for $SUV_{max}$ and $SUV_{peak}$, although it was less significant because their respective increase was not correlated with the MATV. Therefore, PVE correction might have further reduced the clinical value of SUV measurements in this context. This effect has been previously suggested as a limitation to the prognostic value of $SUV_{max}$ in early-stage non–small cell lung cancer (*37*).

Similar conclusions can be drawn from the results regarding the impact of PVE correction on the prognostic value of the SUV parameters. Indeed, as already demonstrated (*12*), extreme MATV values were significantly associated with longer or shorter overall survival for very small (49 mo for MATV below 15 $cm^3$ vs. 11 mo for MATV above 15 $cm^3$) or very large MATV (6 mo for tumor volume above 85 $cm^3$ vs. 20 mo for MATV below 80 $cm^3$), respectively. On the other hand, SUV measurements without correction cannot significantly differentiate between the patients with longer or shorter survival ($P > 0.05$ for all SUV measurements), although a trend for longer survival was associated with lower SUVs. After correction, this differentiation was not significantly improved, because SUVs associated with the smaller volumes were closer to SUVs associated with larger volumes. Therefore, the discrimination was again reduced by PVE correction. To our knowledge there are no similar data available on the impact of PVE correction on SUV predictive value in the literature, but our results are in agreement with previous findings that demonstrated no significant changes in disease-free survival correlation between original and corrected SUVs in esophageal cancer using alternative less accurate methodologies for both PVE correction and functional volume segmentation (*19*).

As previously demonstrated (*7,12*), MATV and associated TLG values were good predictors of response (*7*) and independent prognostic factors of overall survival (*12*). After PVE correction, the already high clinical value of MATV and TLG was not significantly altered. Considering the thresholds used to differentiate patient groups, there was no need for adjustment regarding MATV measurements because MATVs were not significantly modified by PVE correction. On the other hand, TLG thresholds needed to be adjusted, considering that PVE correction led to significantly increased $SUV_{mean}$ and resulting TLG values. The determined threshold values for each parameter regarding prognosis or prediction of response were found using receiver-operating-characteristic analysis on the current patient cohort and would therefore require larger prospective studies to be validated.

The rather large tumor volumes (40 ± 30 $cm^3$) in our patient dataset might be considered as a limitation of this study, because PVEs are usually considered significant for volumes around or below 10 $cm^3$ (*33*). First, 25% of the tumors in this dataset were within this volume range. In addition, the shape of the primary esophageal lesions is not spheric but mostly cylindric, with a small diameter (<2 cm) in the transaxial direction. Therefore, esophageal lesions can be significantly affected by PVEs despite the overall large metabolic volumes, as can be seen in Figure 2 for a lesion with a MATV above 25 $cm^3$. Finally, the patient population used in this study was typical of routine clinical practice and was not selected on the basis of the overall primary MATVs.

## CONCLUSION

The results of this study demonstrate that PVE correction does not add any value to parameters derived from MATVs such as MATV and TLG measured on $^{18}$F-FDG PET baseline acquisitions. PVE correction did not alter the already demonstrated clinical value of both parameters as predictive factors of the response to concomitant radiochemotherapy or as prognostic factors of overall survival in locally advanced esophageal cancer. Similarly, although PVE correction led to increases in all SUV measurements (maximum, peak, or mean) considered in clinical practice, the corrected values were still not significantly associated with either therapy response or prognosis. Finally, our study is in agreement with previous investigations using simpler tools, showing limited interest in PVE correction in this specific context. However, the potential impact of PVE correction in other applications such as diagnosis or lesion detectability remains to be evaluated. In addition, the value of PVE correction in patient follow-up using serial PET scans needs to be further demonstrated.

## DISCLOSURE STATEMENT

The costs of publication of this article were defrayed in part by the payment of page charges. Therefore, and solely to indicate this fact, this article is hereby marked "advertisement" in accordance with 18 USC section 1734.

## ACKNOWLEDGMENT

No potential conflict of interest relevant to this article was reported.

## REFERENCES

1. Parkin DM, Bray F, Ferlay J, Pisani P. Global cancer statistics, 2002. *CA Cancer J Clin.* 2005;55:74–108.
2. Hayat MJ, Howlader N, Reichman ME, Edwards BK. Cancer statistics, trends, and multiple primary cancer analyses from the Surveillance, Epidemiology, and End Results (SEER) Program. *Oncologist.* 2007;12:20–37.
3. Gebski V, Burmeister B, Smithers BM, Foo K, Zalcberg J, Simes J. Survival benefits from neoadjuvant chemoradiotherapy or chemotherapy in oesophageal carcinoma: a meta-analysis. *Lancet Oncol.* 2007;8:226–234.

4. Kelsen DP, Winter KA, Gunderson LL, et al. Long-term results of RTOG trial 8911 (USA Intergroup 113): a random assignment trial comparison of chemotherapy followed by surgery compared with surgery alone for esophageal cancer. *J Clin Oncol.* 2007;25:3719–3725.
5. Stahl M, Wilke H, Stuschke M, et al. Clinical response to induction chemotherapy predicts local control and long-term survival in multimodal treatment of patients with locally advanced esophageal cancer. *J Cancer Res Clin Oncol.* 2005;131:67–72.
6. Aklilu M, Ilson DH. Targeted agents and esophageal cancer: the next step? *Semin Radiat Oncol.* 2007;17:62–69.
7. Hatt M, Visvikis D, Pradier O, Cheze-le Rest C. Baseline $^{18}$F-FDG PET image-derived parameters for therapy response prediction in oesophageal cancer. *Eur J Nucl Med Mol Imaging.* 2011;38:1595–1606.
8. Larson SM, Erdi Y, Akhurst T, et al. Tumor treatment response based on visual and quantitative changes in global tumor glycolysis using PET-FDG imaging: the visual response score and the change in total lesion glycolysis. *Clin Positron Imaging.* 1999;2:159–171.
9. Hyun SH, Choi JY, Shim YM, et al. Prognostic value of metabolic tumor volume measured by $^{18}$F-fluorodeoxyglucose positron emission tomography in patients with esophageal carcinoma. *Ann Surg Oncol.* 2010;17:115–122.
10. Roedl JB, Prabhakar HB, Mueller PR, Colen RR, Blake MA. Prediction of metastatic disease and survival in patients with gastric and gastroesophageal junction tumors: the incremental value of PET-CT over PET and the clinical role of primary tumor volume measurements. *Acad Radiol.* 2009;16:218–226.
11. Mamede M, Abreu ELP, Oliva MR, Nose V, Mamon H, Gerbaudo VH. FDG-PET/CT tumor segmentation-derived indices of metabolic activity to assess response to neoadjuvant therapy and progression-free survival in esophageal cancer: correlation with histopathology results. *Am J Clin Oncol.* 2007;30:377–388.
12. Hatt M, Visvikis D, Albarghach NM, Tixier F, Pradier O, Cheze-le Rest C. Prognostic value of $^{18}$F-FDG PET image-based parameters in oesophageal cancer and impact of tumour delineation methodology. *Eur J Nucl Med Mol Imaging.* 2011;38:1191–1202.
13. Wahl RL, Jacene H, Kasamon Y, Lodge MA. From RECIST to PERCIST: evolving considerations for PET response criteria in solid tumors. *J Nucl Med.* 2009;50(suppl 1):122S–150S.
14. Lee HY, Hyun SH, Lee KS, et al. Volume-based parameter of $^{18}$F-FDG PET/CT in malignant pleural mesothelioma: prediction of therapeutic response and prognostic implications. *Ann Surg Oncol.* 2010;17:2787–2794.
15. Cazaentre T, Morschhauser F, Vermandel M, et al. Pre-therapy $^{18}$F-FDG PET quantitative parameters help in predicting the response to radioimmunotherapy in non-Hodgkin lymphoma. *Eur J Nucl Med Mol Imaging.* 2010;37:494–504.
16. Tixier F, Le Rest CC, Hatt M, et al. Intratumor heterogeneity characterized by textural features on baseline $^{18}$F-FDG PET images predicts response to concomitant radiochemotherapy in esophageal cancer. *J Nucl Med.* 2011;52:369–378.
17. Kwee RM. Prediction of tumor response to neoadjuvant therapy in patients with esophageal cancer with use of $^{18}$F FDG PET: a systematic review. *Radiology.* 2010;254:707–717.
18. Hoetjes NJ, van Velden FH, Hoekstra OS, et al. Partial volume correction strategies for quantitative FDG PET in oncology. *Eur J Nucl Med Mol Imaging.* 2010;37:1679–1687.
19. van Heijl M, Omloo JM, van Berge Henegouwen MI, van Lanschot JJ, Sloof GW, Boellaard R. Influence of ROI definition, partial volume correction and SUV normalization on SUV-survival correlation in oesophageal cancer. *Nucl Med Commun.* 2010;31:652–658.
20. Hatt M, Cheze le Rest C, Descourt P, et al. Accurate automatic delineation of heterogeneous functional volumes in positron emission tomography for oncology applications. *Int J Radiat Oncol Biol Phys.* 2010;77:301–308.
21. Therasse P, Arbuck SG, Eisenhauer EA, et al. New guidelines to evaluate the response to treatment in solid tumors. European Organization for Research and Treatment of Cancer, National Cancer Institute of the United States, National Cancer Institute of Canada. *J Natl Cancer Inst.* 2000;92:205–216.
22. Boussion N, Cheze Le Rest C, Hatt M, Visvikis D. Incorporation of wavelet-based denoising in iterative deconvolution for partial volume correction in whole-body PET imaging. *Eur J Nucl Med Mol Imaging.* 2009;36:1064–1075.
23. Lucy LB. An iteration technique for the rectification of observed distributions. *Astron J.* 1974;79:745–754.
24. Richardson WH. Bayesian-based iterative method of image restoration. *J Opt Soc Am.* 1972;62:55–59.
25. Van Cittert PH. Zum einfluss der spaltbreite auf die intensita tsverteilung in spektrallinien [German]. *Z Physik.* 1931;69:298.
26. Chang SG, Yu B, Vetterli M. Adaptive wavelet thresholding for image denoising and compression. *IEEE Trans Image Process.* 2000;9:1532–1546.
27. Hatt M, Cheze le Rest C, Turzo A, Roux C, Visvikis D. A fuzzy locally adaptive Bayesian segmentation approach for volume determination in PET. *IEEE Trans Med Imaging.* 2009;28:881–893.
28. Hatt M, Cheze Le Rest C, Albarghach N, Pradier O, Visvikis D. PET functional volume delineation: a robustness and repeatability study. *Eur J Nucl Med Mol Imaging.* 2011;38:663–672.
29. Hatt M, Cheze-Le Rest C, Aboagye EO, et al. Reproducibility of $^{18}$F-FDG and 3′-deoxy-3′-$^{18}$F-fluorothymidine PET tumor volume measurements. *J Nucl Med.* 2010;51:1368–1376.
30. Kruskal W, Wallis W. Use of ranks in one-criterion variance analysis. *J Am Stat Assoc.* 1952;47:583–621.
31. Kaplan E, Meyer P. Non parametric estimation from incomplete observation. *J Am Stat Assoc.* 1958;53:457–481.
32. Metz CE. Basic principles of ROC analysis. *Semin Nucl Med.* 1978;8:283–298.
33. Soret M, Bacharach SL, Buvat I. Partial-volume effect in PET tumor imaging. *J Nucl Med.* 2007;48:932–945.
34. Hatt M, Visvikis D. Defining radiotherapy target volumes using $^{18}$F-fluoro-deoxy-glucose positron emission tomography/computed tomography: still a Pandora's box?– in regard to Devic et al. (Int J Radiat Oncol Biol Phys 2010). *Int J Radiat Oncol Biol Phys.* 2010;78:1605.
35. Omloo JM, Sloof GW, Boellaard R, et al. Importance of fluorodeoxyglucose-positron emission tomography (FDG-PET) and endoscopic ultrasonography parameters in predicting survival following surgery for esophageal cancer. *Endoscopy.* 2008;40:464–471.
36. Chatterton BE, Ho Shon I, Baldey A, et al. Positron emission tomography changes management and prognostic stratification in patients with oesophageal cancer: results of a multicentre prospective study. *Eur J Nucl Med Mol Imaging.* 2009;36:354–361.
37. Hanin FX, Lonneux M, Cornet J, et al. Prognostic value of FDG uptake in early stage non-small cell lung cancer. *Eur J Cardiothorac Surg.* 2008;33:819–823.

TB, PNM, PMB/210121, 14/02/2006

INSTITUTE OF PHYSICS PUBLISHING PHYSICS IN MEDICINE AND BIOLOGY

Phys. Med. Biol. **51** (2006) 1–20 UNCORRECTED PROOF

# A multiresolution image based approach for correction of partial volume effects in emission tomography

**N Boussion**[1], **M Hatt, F Lamare, Y Bizais, A Turzo, C Cheze-Le Rest and D Visvikis**

INSERM U650, Laboratoire du Traitement de l'Information Médicale (LaTIM), CHU Morvan, Brest, France

Received 11 October 2005, in final form 23 January 2006
Published DD MMM 2006
Online at stacks.iop.org/PMB/51/1

**Abstract**

Partial volume effects (PVE) are consequences of the limited spatial resolution in emission tomography. They lead to a loss of signal in tissues of size similar to the point spread function and induce activity spillover between regions. Although PVE can be corrected for by using algorithms that provide the correct radioactivity concentration in a series of regions of interest (ROIs), so far little attention has been given to the possibility of creating improved images as a result of PVE correction. Potential advantages of PVE-corrected images include the ability to accurately delineate functional volumes as well as improving tumour-to-background ratio, resulting in an associated improvement in the analysis of response to therapy studies and diagnostic examinations, respectively. The objective of our study was therefore to develop a methodology for PVE correction not only to enable the accurate recuperation of activity concentrations, but also to generate PVE-corrected images. In the multiresolution analysis that we define here, details of a high-resolution image H (MRI or CT) are extracted, transformed and integrated in a low-resolution image L (PET or SPECT). A discrete wavelet transform of both H and L images is performed by using the 'à trous' algorithm, which allows the spatial frequencies (details, edges, textures) to be obtained easily at a level of resolution common to H and L. A model is then inferred to build the lacking details of L from the high-frequency details in H. The process was successfully tested on synthetic and simulated data, proving the ability to obtain accurately corrected images. Quantitative PVE correction was found to be comparable with a method considered as a reference but limited to ROI analyses. Visual improvement and quantitative correction were also obtained in two examples of clinical images, the first using a combined PET/CT scanner with a lymphoma patient and the second using a FDG brain PET and corresponding T1-weighted MRI in an epileptic patient.

[1] Author to whom any correspondence should be addressed.

## 1. Introduction

Partial volume effects (PVE) are well-known consequences of the limited spatial resolution in emission tomography. PVE is characterized by the loss of signal in tissues of size similar to the point spread function (PSF). In addition, PVE induces a signal cross-contamination in adjacent structures with different amounts of radioactivity (Aston *et al* 2002, Du *et al* 2005). In this latter phenomenon, sometimes referred to as spillover, the high activity in a given region can spread out and contaminate a bordering area of lower activity, leading to either underestimated or overestimated activity concentration measurements.

These effects can be corrected for by using a number of different algorithms that often rely on the use of the PSF of the imaging device and *a priori* anatomical knowledge provided by computed tomography (CT) or magnetic resonance imaging (MRI) (Meltzer *et al* 1990, Muller-Gartner *et al* 1992, Rousset *et al* 2000, Aston *et al* 2002, Matsuda *et al* 2003, Baete *et al* 2004, Bencherif *et al* 2004, Quarantelli *et al* 2004, Kusano *et al* 2005, Rota Kops and Krause 2005). The large majority of these algorithms, which have been evaluated mostly in the context of cerebral imaging, require a segmentation step to delineate the different parts from anatomical images. This step renders their accuracy dependent on the segmentation algorithm used as well as making their application on other clinical investigations outside the brain challenging (Feuardent *et al* 2003). For example, the pixel-based approach of Meltzer *et al* (1990) is restricted indeed to brain metabolism or neuroreceptor binding, and requires compartmental analysis (Meltzer *et al* 1999). As a rare example, Pretorius and King (2004) proposed an application of PVE correction for cardiac SPECT. Furthermore, and similar to the great majority of PVE correction methods (except in the interesting approach of Kennedy *et al* using Taylor expansion (Kennedy *et al* 2005)), these algorithms offer quantitative correction of ROI (region of interest) intensities without considering the construction of enhanced images. On the other hand, resolution compensation or resolution recovery algorithms can also be used to reduce PVE in emission tomography. However, the majority of these algorithms suffer from being reconstruction algorithm specific (Ardekani *et al* 1996, Som *et al* 1998, Somayajula *et al* 2005), as well as being only tested in limited clinical context such as cardiac SPECT (Hutton and Lau 1998) or FDG PET in the human brain (Baete *et al* 2004).

One of the reference methods (referred to from here onwards as RSF for regional spread function) described by Rousset *et al* (1998, 2000) and recently improved (Frouin *et al* 2002, Du *et al* 2005) was also developed in the brain context and allows estimating the true mean signal in any user-defined series of $n$ homogeneous regions of interest (ROIs), but the images themselves are not enhanced. This approach relies on the inversion of an $n \times n$ matrix called geometric transfer matrix (GTM). The elements $w_{ij}$ of the GTM are the coefficients of activity spillage from ROI $i$ to ROI $j$, and the true activity $T_i$ in ROI $i$ can be deduced from the measured activity $t_i$ by inverting the equation $[t] = [\mathrm{GTM}] \times [T]$, where $[t]$ and $[T]$ are the vectors containing the $t_i$ and $T_i$ values, respectively. The use of this approach is theoretically possible in various clinical applications even if it was originally designed for cerebral studies where generally only three ROIs are required (white matter, grey matter, cerebrospinal fluid). Actually, the method works satisfactorily when the image is segmented into a series of ROIs that constitute a partition. In other words, ROIs must not overlap and at the same time considering all ROIs together must cover the entire image. As a consequence, when studying tumours in whole-body images, the number of ROIs can dramatically increase (Feuardent *et al* 2003), thus hampering the clinical use of this methodology.

In general, the aim of all these methods is to provide the user with correct radioactivity concentration estimates in a given ROI. To date however, except in very specific applications (Baete *et al* 2004), little attention has been given to the challenging possibility of creating

improved images through a generic approach. In this paper, a new PVE correction methodology is proposed, based on the multiresolution analysis of images of different spatial resolutions. The main advantage of the proposed methodology is that it not only enables the accurate recuperation of activity concentrations, but is also capable of simultaneously generating PVE-corrected images. This improvement allows (a) performing a visual control of the correction, (b) improving clinical diagnostic studies through a better visual assessment of the images, and most important (c) allowing further image processing (such as, for example, functional volume estimate of tumours, location of epiletogenic foci in cerebral imaging (Boussion *et al* 2003), or wall motion and ejection fraction in cardiac imaging (Hickey *et al* 2004)). Furthermore, the method is not restricted to a particular organ and does not require tedious and time-consuming ROI delineation.

In the following section, a concise presentation of the wavelet transform and the multiresolution analysis serves as an introduction to section 2. The developed PVE algorithm is described in detail including a description of the test images and overall methodology used to validate the developed algorithm.

## 2. Materials and methods

### *2.1. Multiresolution image analysis and wavelet transform*

Although the theoretical foundations of multiresolution analysis do not constitute the main topic of this study, it is constructive to introduce the basic concepts of the wavelet transform which is an important part of the proposed methodology. Actually, the wavelet transform can be introduced by comparison with the more common Fourier transform with which it has a number of similarities. While the Fourier transform provides global information about the spatial frequencies in an image, the wavelet transform leads to a local representation of these spectral properties. From an image processing point of view, the Fourier transform permits one to switch between the spatial and the frequency domains while the wavelet transform allows one to bring them together in one single image. In practice, the wavelet transform of a given image is another image presenting the areas where one may find either more or less important contrast. In addition, one of the interests of the wavelet transform in image processing is that it enables work at different levels of spatial resolution, operating as a tool of multiresolution analysis. Multiresolution analysis allows retrieving the layers of details that have different sizes by separating the spatial frequencies that the image contains. Basically, a medical image at a given spatial resolution $R$ contains information at different scales, from large structures to small details. For instance, in a cerebral MRI the sharp edges between white and grey matters will be lost when a low-pass filter is applied, but at the same time the skull will stay clearly separated from the brain. Accessing and separating these structures of different sizes is the scope of multiresolution analysis.

If we now consider the mathematic point of view, the wavelet transform allows expressing a signal according to a basis of elementary functions called wavelets. This basis is built from a 'mother' wavelet $\psi$ (also referred to as analysing wavelet) on which are applied dilation and translation computations. This process is obtained in one dimension as a result of the following formula:

$$\psi_{a,b}(x) = \frac{1}{\sqrt{a}}\psi\left(\frac{x-b}{a}\right) \qquad (a > 0). \tag{1}$$

$a$ is called the scale parameter and is linked to the frequency domain, while $b$ is the position parameter linked to time or space.

The wavelet transform $W(a, b)$ of the function $f(x)$ is defined as

$$W(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{+\infty} f(x)\psi^* \left( \frac{x-b}{a} \right) \mathrm{d}x, \tag{2}$$

where $\psi^*$ stands for the complex conjugate of the analysing wavelet $\psi$. $W$ is linear, shift invariant and also invariant by dilation. These latter two properties are of interest in image processing involving combination of different images. Actually, on the one hand, shift invariance limits the unavoidable consequences of inaccurate superimposition of images. On the other hand, dilation invariance is valuable for observing 'objects' of different sizes in a given signal without changing the analysing wavelet.

There are many algorithms available to perform the discrete wavelet transform of an image. All have particular interests and drawbacks but they must be chosen carefully because the passage to the discrete approach can lead to the loss of interesting properties such as invariance mentioned above. A widely used approach is the pyramidal methodology which consists of reducing the size of the image iteratively to get smoother and smoother versions of the initial image. This is the widespread multiresolution approach that Mallat (1989) developed through his algorithm that permits compression of data by decimating the image. This method is anisotropic in the sense that horizontal, diagonal and vertical details are separated during the process. Another common approach is the algorithm 'à trous' (French term that means 'with holes'). This is an undecimated method inducing shift invariance which is of particular interest when investigating image comparison. The transformation is not pyramidal since the initial image and the images of coarser spatial resolution have identical sizes. For this reason, this particular algorithm is redundant and is of reduced interest in image compression. This algorithm forms however the basis of our PVE correction methodology as it presents several practical advantages, namely (a) the implementation is straightforward and the initial image can be perfectly reconstructed without loss of any kind, (b) there is no selection of specific directions during the analysis since the process is isotropic, (c) the transform is known for each pixel improving accuracy of further processing, and (d) navigation is easy between the different levels of resolution.

This discrete wavelet transform algorithm called 'à trous' was introduced by Dutilleux (1987), developed by Holdschneider *et al* (1989) and detailed by Starck *et al* (1998). The process gives an image sequence of coarser and coarser spatial resolution by performing successive convolutions with a low-pass filter $h$ obtained from a scaling function $\phi$. At each iteration $j$, the spatial resolution of the image $I_j$ is degraded to give the approximation image $I_{j+1}$ according to

$$I_{j+1}(k,l) = \sum_{m,n} h(m,n) I_j(k + m2^j, l + n2^j). \tag{3}$$

As already pointed out, there is no decimation involved in the process, which means that all $I_j$ approximations have the size of the initial image $I_0$. However, only one pixel out of $2^j$ is considered during the filtering process, leading to inclusion of zeros in the rows and the columns of the mask. This feature gives its name to the algorithm, i.e. 'with holes', and it also explains why the process is dyadic, where the successive approximations $I_j$ have resolutions decreasing by powers of 2.

The difference $I_j - I_{j+1}$ is the wavelet coefficients $w_{j+1}$ containing the details (edges, texture) at a resolution level between $I_j$ and $I_{j+1}$. Note that the undecimation permits one to follow the local information at a pixel level for any $I_j$, that is, navigation through all $I_j$ images

is possible at any pixel location. The synthesis procedure that reconstructs the original image from its layers of details $w_k$ is given by

$$I_0 = I_N + \sum_{k=1}^{k=N} w_k, \tag{4}$$

with $N$ the number of iterations from the initial image $I_0$ to the final approximation $I_N$ of spatial resolution decreased by $2^N$. A pixel at location $(x, y)$ can be expressed as the sum of the wavelet coefficients at this position plus the smoothed array at the same $(x, y)$ coordinates:

$$I_0(x, y) = I_N(x, y) + \sum_{k=1}^{k=N} w_k(x, y). \tag{5}$$

The 'à trous' algorithm can easily be implemented by performing the following steps (Starck *et al* 1998):

(1) Initialize $j$ to 0: start with the original image $I_0$.
(2) Increment $j$ and carry out a convolution of $I_{j-1}$ with the low-pass filter $h$. The distance between the central pixel and the adjacent ones is $2^{j-1}$.
(3) The wavelet coefficients $w_j$ at this level of resolution are given by $I_{j-1} - I_j$.
(4) If $j$ is less than the required number $N$ of resolutions, go to step 2.
(5) The set $W = \{w_1, w_2, \ldots, w_N, I_N\}$ is the wavelet transform of $I_0$.

Provided they satisfy a limited number of properties (compacity, regularity, symmetry) and according to suitable prerequisites, different scaling functions can be constructed. However, several already exist possessing interesting characteristics. The most widely used filters in the 'à trous' algorithm are based on linear interpolation and B-splines interpolation. For instance, the bicubic spline is a very smooth function, well suited for isolation of large image structures. On the other hand, linear interpolation is a good compromise, enabling work with both small and large scale characteristics. Another filter, sometimes called low-scale filter, is a sharply peaked function that performs well in isolating very small structures. The normalized coefficients of these different filters are presented in the appendix. Each of these filters were tested under different image characteristics in order to evaluate their behaviour and choose the most appropriate one in the framework of the developed PVE correction algorithm.

### *2.2. Description of the algorithm implementation*

The process employed in the developed algorithm comes from the field of data fusion (Luo *et al* 2002) and as stated above it relies on a wavelet-based image merging. Actually, new approaches to image merging that uses multiresolution analysis procedures based upon the discrete wavelet transform have been proposed recently in as different domains as texture classification (Li and Shawe-Taylor 2005), forensic science (Wen and Chen 2004) or aerial images (Ranchin and Wald 2000). The multiscale fusion that we define here is the process whereby details of a high-resolution image H (MRI or CT typically) are extracted, transformed according to a given model and integrated in a low-resolution image L like PET or SPECT for instance. The challenge is to preserve the global functional characteristic of L while incorporating additional data in it and the mandatory hypothesis is that the tissues examined by L are also present in the high-resolution image H. Contrary to all other applications that have been studied till now, such as aerial imaging or forensic sciences, the visual enhancement is not here the unique goal of the process. Our primary objective is the quantitative improvement of the recovered activity concentrations. This latter is possible by adding detail layers of different resolutions that all have a zero-mean signal.

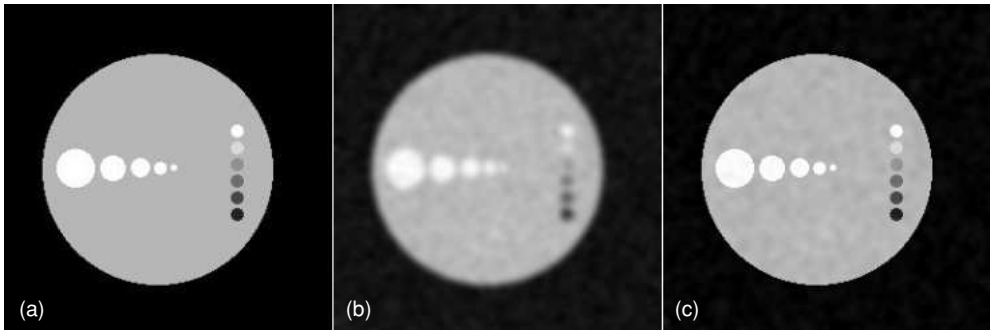



**Figure 1.** (a) The high-resolution H image with discs of various sizes and intensities. The five horizontal discs have identical intensities and decreasing sizes in order to evaluate the correction of tissue-fraction effects. The six vertical discs have decreasing intensities to allow studying spillover effects. (b) Low-resolution image L corresponding to the degradation of high-resolution image H after 10-standard-deviation Gaussian noise addition and low-pass filtering. (c) L image after PVE correction using the developed wavelet-based algorithm.

Wavelet analysis allows the spatial frequencies to be obtained easily at any level of resolution, in particular at a level of resolution common to H and L. A model is then inferred to compute the lacking details of L from the high-frequency details' layers of H. If the level of resolution of H is $q$, referred to as $H_q$, and that of L is $r = q + p$, referred to as $L_r$, we can write

$$L_r(x, y) = L_{q+p}(x, y) = L_{q+p+1}(x, y) + w^{\mathrm{L}}_{q+p+1}(x, y) \tag{6}$$

and

$$H_q(x, y) = H_{q+p+1}(x, y) + \sum_{k=1}^{k=p+1} w^{\mathrm{H}}_{q+k}(x, y). \tag{7}$$

The lacking details of L are the wavelet coefficients $w^{\mathrm{L}}_i$ with $q \leqslant i \leqslant q + p$. However, we do possess $w^{\mathrm{L}}_{q+p+1}$ and $w^{\mathrm{H}}_{q+p+1}$ and we assume that there exists a more or less simple link between them like $w^{\mathrm{L}}_{q+p+1} = \alpha \times w^{\mathrm{H}}_{q+p+1}$, $\alpha \in IR^*$ for instance. Although, different models can be envisaged, in this study a simple linear model is used where the parameter $\alpha$ is considered equal to the mean pixel-by-pixel division of $w^{\mathrm{L}}_{q+p+1}$ by $w^{\mathrm{H}}_{q+p+1}$.

$L_q$ can now be reconstructed from $L_r$ by taking $w^{\mathrm{L}}_i$ ($q \leqslant i \leqslant q + p$) into account. They are calculated as $w^{\mathrm{L}}_i = \alpha \times w^{\mathrm{H}}_i$ ($q \leqslant i \leqslant q + p$) leading to

$$L_q(x, y) = L_{q+p+1}(x, y) + \alpha \sum_{k=1}^{k=p+1} w^{\mathrm{H}}_{q+k}. \tag{8}$$

### *2.3. Validation studies*

*2.3.1. Synthetic and simulated images.* The developed algorithm was firstly validated using different synthetic and simulated datasets. Synthetic images were composed of a circular container (intensity 50) including discs of different sizes and contrast ratios (figure 1(a)). A first series of five horizontal discs of decreasing diameter (30 mm, 20 mm, 15 mm, 10 mm, 5 mm) and constant intensity (70) was built up to specifically examine the tissue-fraction effect and the recovering of small areas. A second series of six vertical discs of constant diameter (10 mm) but decreasing intensity (70, 60, 40, 30, 20 and 10) was designed specifically to consider

spillover effects. This image of high spatial resolution (1 mm) corresponded to H as stated in the previous section. A *n*-standard-deviation (SD) Gaussian noise was added (*n* ranging from 1 to 10) and a 6 mm FWHM Gaussian blur was convolved in order to simulate the L images (figure 1(b)) with a uniform 6 mm spatial resolution (the letter *n* in *n*-standard deviation corresponds to the amount of noise added on a given pixel; the Gaussian law built with a zero mean and a standard deviation equal to *n* times the standard deviation calculated in a 5 $\times$ 5 pixel ROI around the pixel to treat). The convolution induced a contamination of signal between homogeneous areas of the synthetic images very similar to partial volume effects, and the levels of resolution for L and H were chosen according to those of typical PET and MRI studies, respectively. The different amounts of noise in L, introduced by the variable SD Gaussian noise, aimed at investigating the performance of the correction algorithm in PET images of variable statistical quality.

The mean intensity in the discs inside the container was calculated before and after PVE correction in each one of the ten L images and then compared with actual values in H. In practice, ROIs delineating discs in L were obtained automatically by copying the masks of exact discs in H. The mean intensity in the discs of L images was then calculated inside these exact ROIs. These results were also compared with those obtained by applying the RSF method (Rousset *et al* 1998) often considered as the reference numerical approach (Frouin *et al* 2002, Quarantelli *et al* 2004) for PVE correction in emission tomography.

The robustness of the developed algorithm to spatial misalignment between the H and L images was also studied by introducing artificial displacements of L with regards to H. A set of 20 configurations was created, namely 1, 2 and 3 pixels (each pixel 1 mm of size) translation errors in the four directions (up, down, left, right), 1, 2 and 3 degrees of rotations clockwise and anticlockwise, and finally, two scaling errors of 99% and 101%. The L image with 6 SD Gaussian noise was considered for this specific investigation. The error in intensity recovery was calculated in the 11 discs for each of the 20 configurations of misalignment produced, after applying either the RSF PVE correction or our wavelet-based method.

The synthetic images were also used to assess the proposed methodology in cases where image contents no longer correlate. As already mentioned, one of the mandatory prerequisites to the application of the correction method proposed in this paper is the similarity of tissues in the images we are dealing with. For this purpose, and without altering the low-resolution image, we modified the synthetic 'CT' image to create two grossly unfavourable configurations. In the first one, the horizontal series of five discs was completely removed from the synthetic 'CT' image (figure 2(a)), and in the second one (figure 2(c)) the intensity of the first disc in this same series was set to 20 (cold intensity) instead of 70 (hot intensity, as in the 'PET' image).

Finally, simulated images were also included in our study. They consisted of a simplified numerical version (figure 3(a)) of the physical IEC phantom (IEC Publication 61675-1 1998). This phantom consists of a 20 cm diameter by 20 cm long cylinder, containing six spheres of 37 mm, 28 mm, 22 mm, 17 mm, 13 mm and 10 mm in diameter. The numerical version of this phantom was produced as a set of 64 contiguous planes of 64 $\times$ 64 square pixels of 4 mm $\times$ 4 mm in size. This phantom was subsequently combined with a Monte Carlo based simulation of the Philips Allegro PET scanner using GATE (Lamare *et al* 2006). A total of 60 million coincidences were simulated considering a sphere/cylinder activity concentration ratio of 5/1. Images were subsequently reconstructed using the OPLEM algorithm (11 iterations) (Reader *et al* 2002). The high-resolution image serving for PVE correction was the numerical phantom in which values were arbitrarily set to 1000 for the background, 2000 for the cylindrical container and 3000 for the spheres, leading to a 1.5 sphere/cylinder intensity ratio. These realistic ratios in the numerical phantom and in the simulated PET image were chosen in order

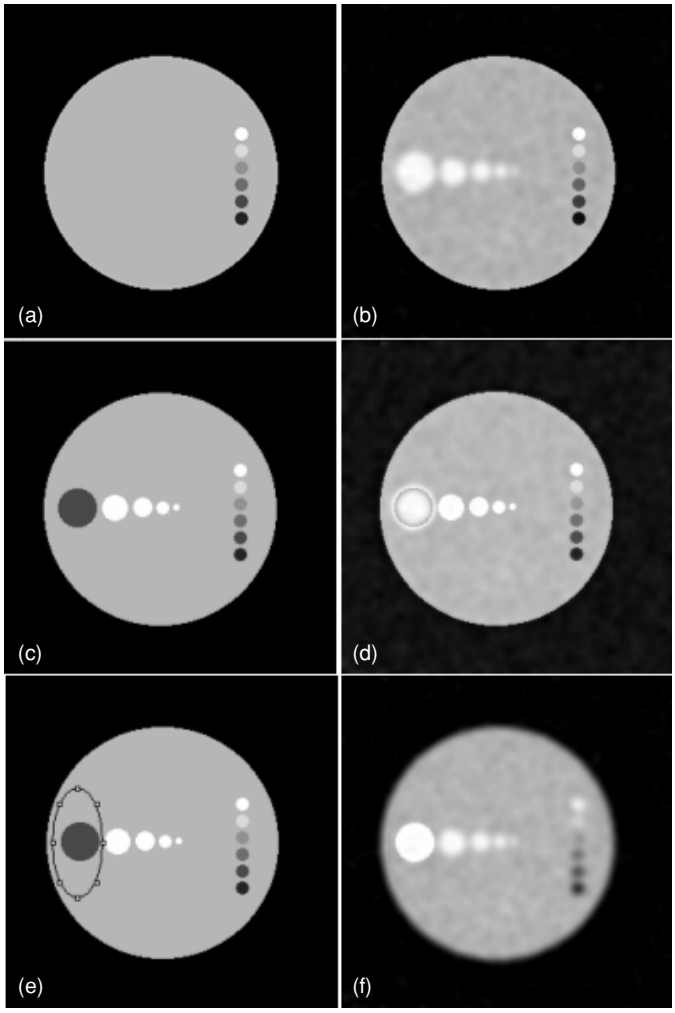



**Figure 2.** Synthetic images with no tissue correlation between high- and low-resolution images; only the high-resolution image is altered. (a) First configuration, without horizontal discs in the high-resolution image, and (b) corresponding PVE-corrected low-resolution image. (c) Second configuration, with contrast modified in the first horizontal disc only, and (d) corresponding PVE-corrected low-resolution image. A local application of the proposed algorithm is shown in (e), with the region of interest surrounding the disc to be corrected, and (f) the whole image after PVE correction demonstrating that only the part inside the specified region of interest is PVE corrected.

to investigate the behaviour of the developed PVE correction methodology in more realistic conditions than the synthetic images.

*2.3.2. Quantitative and qualitative assessment of the developed methodology on clinical images.* In order to demonstrate the use of the developed algorithm in the clinical context, the technique was applied on two different sets of patients' images. The first one consisted of a whole-body FDG PET and corresponding CT images figure 4 acquired on a lymphoma patient using a dedicated combined PET/CT scanner (GE Discovery LS), while the second dataset consisted of a FDG brain PET and corresponding T1-weighted MRI images acquired during pre-surgical evaluation of refractory epilepsy figure 5. The MRI and PET cerebral

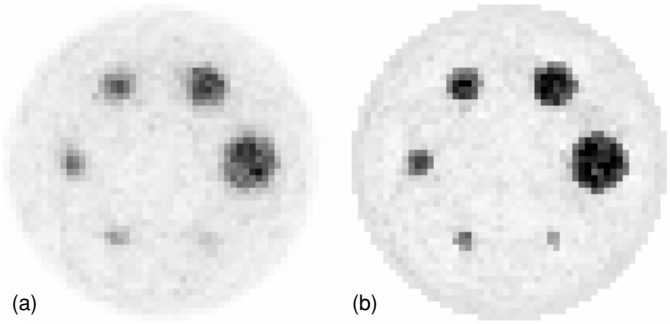



**Figure 3.** Simulated numerical image of the IEC image quality phantom: (a) uncorrected and (b) PVE corrected.

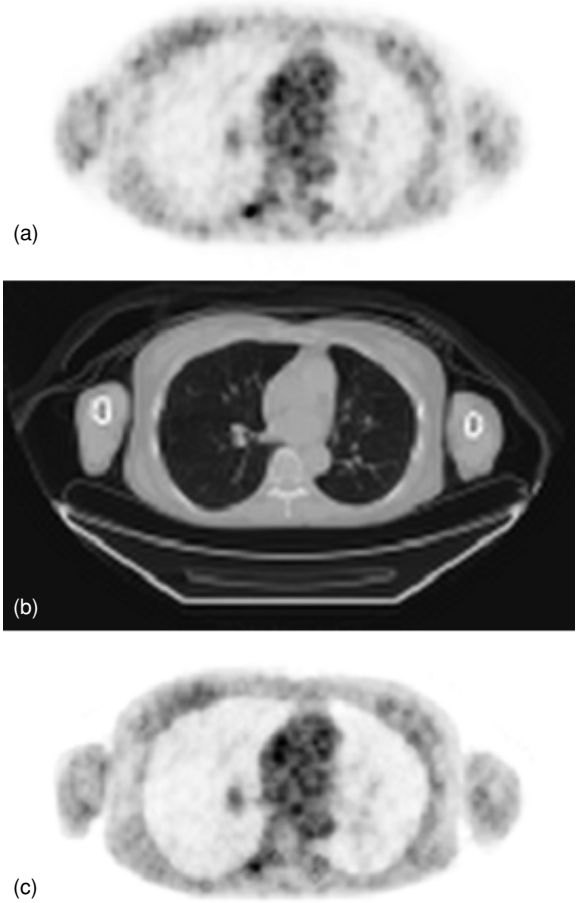



**Figure 4.** Clinical PET/CT patient study of the thorax acquired on a combined PET/CT scanner. (a) Original emission PET FDG image. (b) Corresponding CT image at the same anatomical location (identical slice level). (c) FDG PET after PVE correction.

images were acquired separately and spatially co-registered by using mutual information maximization (Wells *et al* 1996) and affine transformation (rotation, translation, scaling).

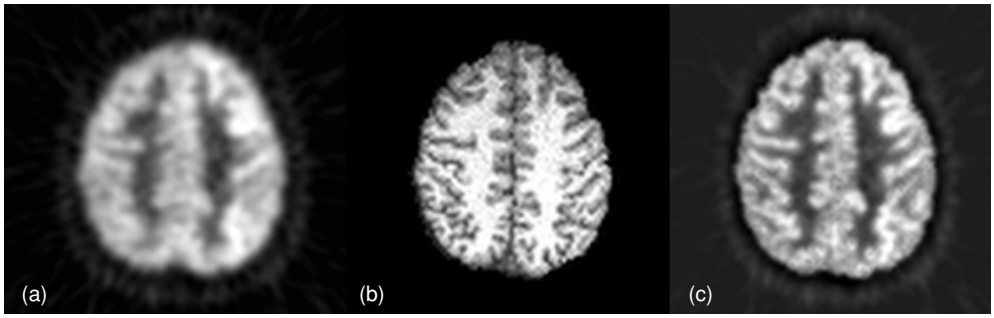



**Figure 5.** (a) Transaxial brain PET image obtained with FDG in an epilepsy follow-up study. (b) Corresponding MRI slice (the skull has been removed). (c) Same transaxial plane as in (a), here shown after PVE correction by using the developed wavelet-based multiresolution analysis algorithm.

Apart from the qualitative assessment of the corrected images, a ROI analysis was also performed to quantify the impact of the PVE correction. Different regions were drawn in both the original and corrected whole-body PET images, namely four circular ROIs of 30 mm in diameter placed at the middle of the right lung corresponding to a normal area with homogeneous intensity ($ROI_{lung}$), two circular ROIs of 20 mm in diameter inside the heart in a normal but visually inhomogeneous area ($ROI_{heart}$) and a ROI surrounding the lesion in left lung ($ROI_{lesion}$, size 20 mm $\times$ 8 mm). A similar quantitative investigation was performed in the FDG brain PET, in which grey matter and white matter intensities were calculated before and after PVE correction. As a first step, grey and white matters were delineated automatically in the T1-weighted MRI using the SPM software (Ashburner and Friston 2000), and in a second step the two obtained segmented areas were superimposed on the PET image. They both served as ROIs in which mean intensities were calculated.

## 3. Results

Figure 1(b) shows an example of the synthetic image L used to assess the developed methodology. The corresponding PVE-corrected image of L is given in figure 1(c), where it can be noted that, aside from quantitative considerations, edges are visually enhanced. Figure 6 illustrates the PVE correction in a semi-quantitative fashion. Profiles across the five horizontal discs in uncorrected and corrected L images along with the related wavelet coefficients built from the H image and representing the correction values are presented. In the profile presented in figure 6(a), one can note that the discs are represented by five coarse Gaussian shapes, the last being so attenuated that it is hardly visible. The corresponding profile in figure 6(b) shows that the correction required for this disc is the most significant among the five discs, leading at the end to equal corrected values as demonstrated by figure 6(c). As a consequence, the process does not only enhance the edges of objects but also increases the global intensity level when needed, especially for smaller objects.

The global recovery of intensity in the low-resolution synthetic images, considering the ten different levels of image noise described in section 2.3.1, is represented in figure 7, where results concerning tissue-fraction and spillover effects are separated in figures 7(a) and 7(b), respectively. As figure 7(a) demonstrates, the intensity level in the five discs with diminishing sizes in the uncorrected L images was found to dramatically decrease clearly demonstrating the tissue-fraction effects. For example, the intensity in disc 1 (30 mm diameter) and disc 5

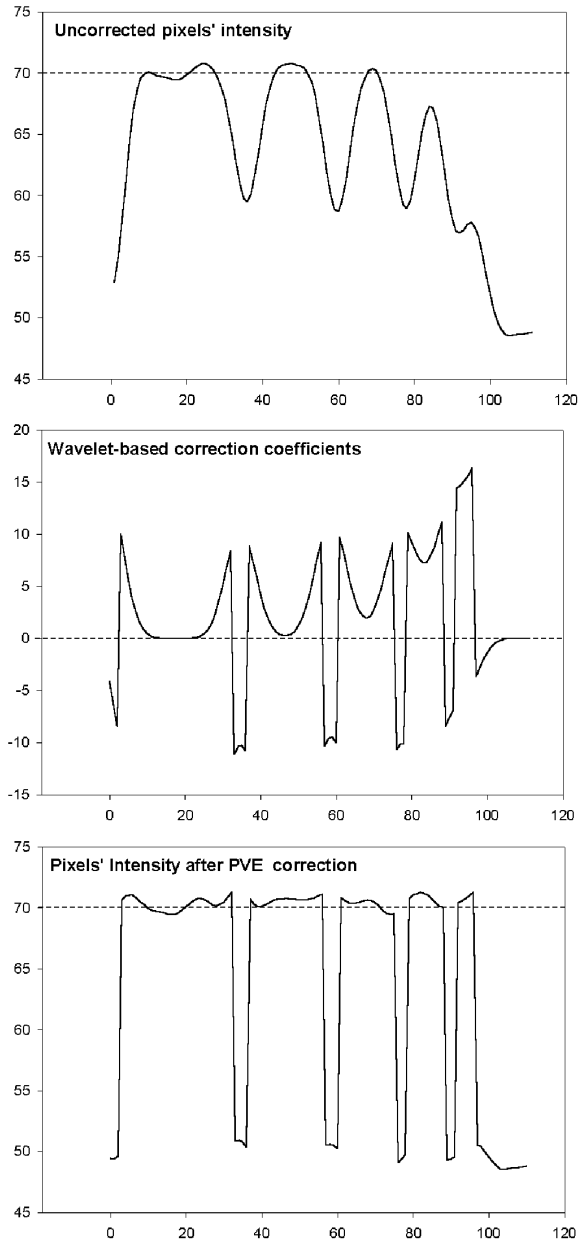



**Figure 6.** A semi-quantitative assessment of the results obtained from the application of the developed PVE methodology on the synthetic images (considering the L image with 5-standard-deviation Gaussian noise). A 'plot profile' is generated along a line crossing the five horizontal discs, in the uncorrected image (a), in the wavelet-based correction image (b) and in the corrected image which is the pixel-to-pixel addition of the first two (c).

(5 mm diameter) decreased from 70 to 67.2 (−4.0%) and 56.7 (−19.0%), respectively. The spillover effect is illustrated by the six vertical discs, whose intensities were either underestimated or overestimated depending on their initial signal-to-background (S/B) ratios (figure 7(b)). For example, the intensity in disc 6 (10 mm in diameter, initial S/B = 1.4)

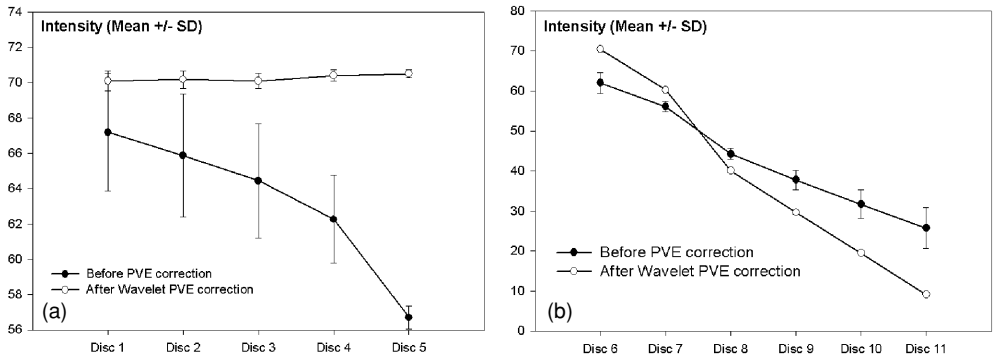



**Figure 7.** Quantitative performance assessment of the PVE correction. The results from the L synthetic images of the ten different noise levels considered are summarized in this figure, namely (a) mean intensity recovery in the five discs of diminishing sizes, with the errors bars representing the standard deviation in the mean taking into account the ten different noise levels considered (the actual intensity is 70), (b) mean intensity in the six discs of decreasing intensities (the actual values are 70, 60, 40, 30, 20 and 10 for discs 6–11, respectively).

decreased by 11.4% and the intensity in disc 11 (10 mm in diameter, initial S/B = 0.2) increased by 158.0% (from 10 to 25.8). Following the application of the developed algorithm, both phenomena were corrected for as also illustrated in figure 7. It is however of significance to observe the dependence of the correction upon the choice of the filter. In figure 8, for example, the results of the correction are shown for the 11 discs simulating the tissue-fraction effect and spillover, according to the four filters described in section 2.1 and the appendix. For this purpose, we call recovery error the difference between the expected disc intensity (for example, 70 in the discs 1–5) and the intensity in the same disc after PVE correction. A perfect PVE correction would then lead to a recovery error equal to 0. It is clear that the bicubic spline filter and the 5 × 5 linear filter perform better than the other two. For discs 1–5, the percentage of recovery error was less than 1% for these two filters, while the low-scale filter led to error greater than 4%. For discs 9–11, the percentage of recovery error exceeded 20% with this low-scale filter. According to these results, the bicubic spline filter or the 5 × 5 linear filter should be preferred. Consequently, all results presented in this paper were obtained with the bicubic spline filter.

The good noise characteristics of the developed algorithm are shown in figure 9 where the percentage of correction error is plotted against the different L images considering variable noise levels. As far as tissue-fraction effects are considered (figure 9(a)), a negligible overestimation error was found, globally increasing with respect to the amount of noise. However, errors never exceeded 1% in the set of images considered. Concerning the spillover, the correction slightly underestimated the true values, with an error up to 4% in high noise conditions figure 9(b). Finally, the comparison with the RSF method is given in figure 10. The graphs show that the two methods have very similar behaviours in correcting for both tissue-fraction effects (discs 1–5) and spillover (discs 6–11) since the two graphic representations almost perfectly overlap. However, a closer comparison reveals a slight difference in favour of the wavelet method. For the discs 4 and 5, which are highly subjected to tissue-fraction effect on account of their small sizes, the errors in intensity recovery are, respectively, 0.07% and 0.33% for the wavelet approach against 0.34% and 0.92% for the RSF method. The difference is more significant when considering the disc 11 which undergoes substantial spill-in from the surrounding area: 0.47% of error for the proposed method against 4.27% for the RSF approach.

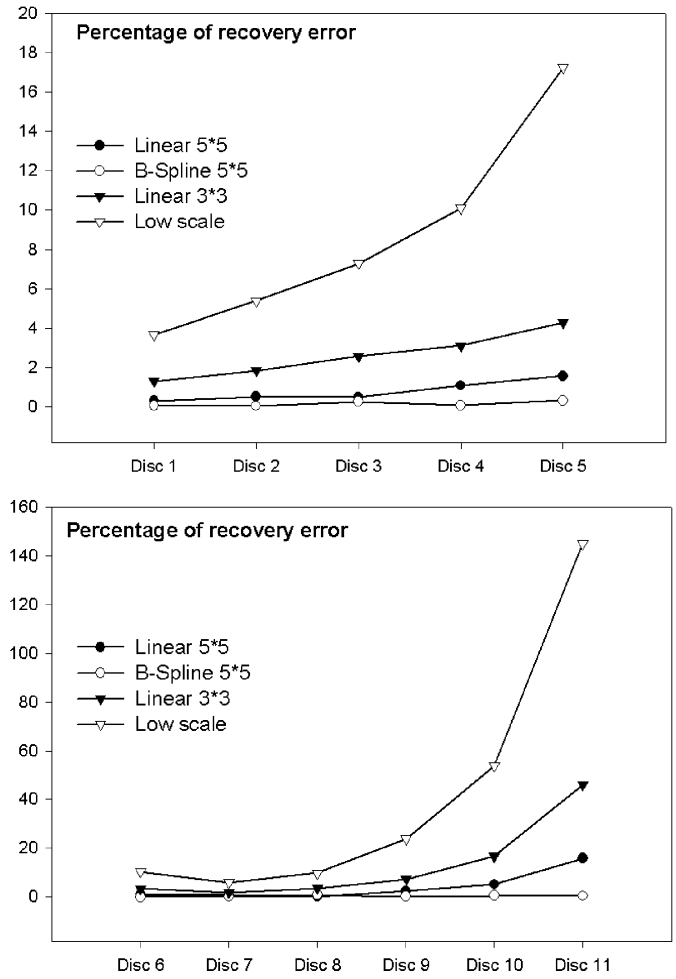



**Figure 8.** Influence of the filter used in the 'à trous' algorithm on the percentage of recovery error. Results are presented for the 11 discs of the synthetic images, dedicated to the study of tissue-fraction effects (discs 1–5) and spillover (discs 6–11). Values are mean percentages obtained from all ten different noise level synthetic images.

Figure 11 summarizes the results on the effects of spatial registration errors, demonstrating that the developed algorithm behaves similarly to the RSF method. Both methods gave satisfactory PVE correction of tissue-fraction effects (discs 1–5). The correction of PVE in discs simulating spillover seemed to be more dependant upon co-registration than the correction of tissue-fraction effects (discs 6–11). However, as figure 11 demonstrates the developed algorithm performed better in correcting spillover effects than the RSF method, since, in some cases, the latter led to errors exceeding 100% for discs 10 and 11.

The effects of grossly unfavourable configurations for the methodology described in this work, where tissues greatly differ between high- and low-resolution images, are demonstrated in figure 2. The image in figure 2(a) corresponds to the case where horizontal discs are erased in the high-resolution image. As a result, in the corresponding corrected image figure 2(b) vertical discs are PVE corrected as is the contour of the container, but the horizontal series of discs stays unmodified since no corresponding information exists in the high-resolution

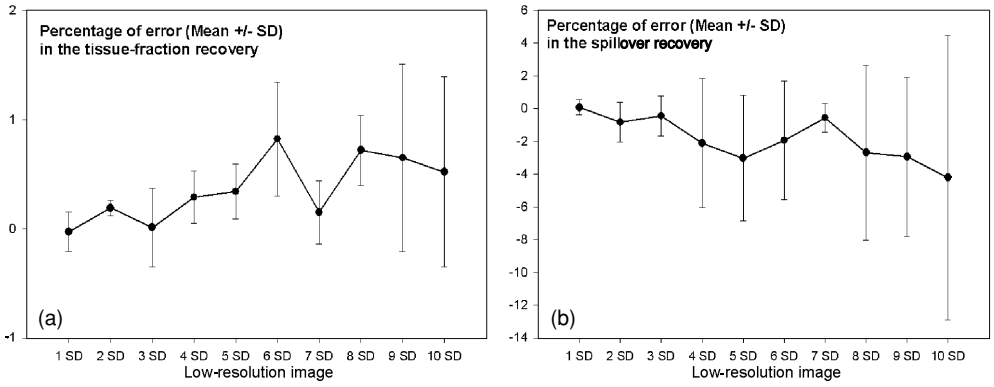



**Figure 9.** Robustness of the wavelet-based PVE correction according to noise. (a) Error percentage of intensity recovery in the five discs of decreasing sizes, and (b) error percentage of intensity recovery in the six discs of decreasing intensities.

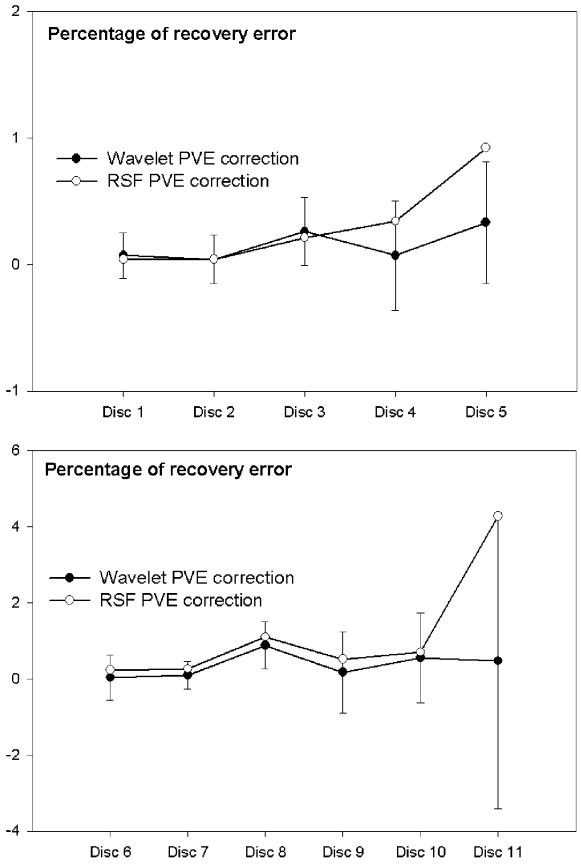



**Figure 10.** Comparison of the wavelet-based PVE correction with the RSF approach of Rousset *et al*, presented in terms of percentage of recovery error. Results are mean values obtained in the set of ten L images with different levels of noise. For the sake of clarity, the standard deviations (error bars) are presented for the wavelet-based method only.

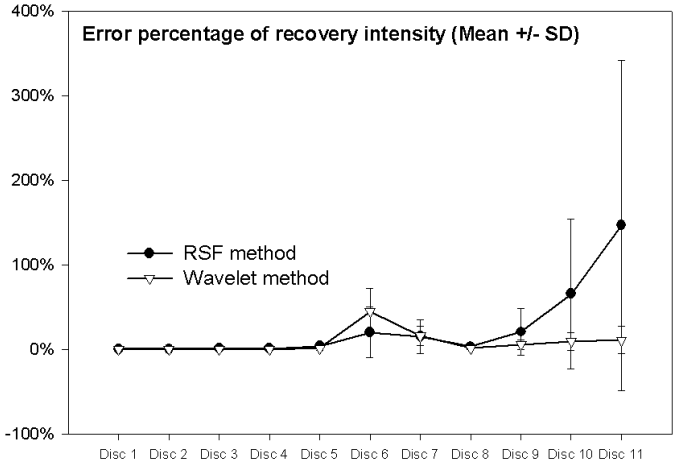



**Figure 11.** Robustness of the correction according to the alignment accuracy of the images. Twenty different configurations were tested (12 translation movements, 6 rotation movements and 2 inadequate scalings). For each disc, the value presented is the mean error of intensity recovery obtained in the set of 20 configurations considered.

image. In terms of quantitative accuracy, there has been no alteration in the values of the horizontal spheres independently of their sizes. Therefore, the complete lack of a given tissue in the high-resolution image does not modify the corresponding part in the low-resolution image. The image in figure 2(c) relates to the second case where the intensity in an isolated disc is altered compared to the other discs ('cold spot' rather than a 'hot spot'). This leads to a local artefact in the corrected image corresponding to the limits of the 'cold sphere' in the high-resolution image, mainly due to a local inverse contrast in the high-resolution image compared to the same area in the low-resolution image (figure 2(d)). The quantitative errors are also local to the visual artefact with values not altered in the rest of the sphere.

The transaxial slice of the reconstructed PET image of the numerical IEC phantom containing all the lesions is shown in figure 3(a). The corresponding PVE-corrected image is given in figure 3(b), while the quantitative results (an expected ratio between spheres and background of 5/1, irrespective of lesion size) are presented in figure 12. As this latter figure demonstrates, although both correction methods lead to an improvement in the sphere/cylinder ratios, the wavelet-based correction performed better in all spheres, particularly for the smallest ones. For example, the ratios for the 13 mm and 10 mm diameter spheres were improved from 2.19 to 5.93 and from 2.03 to 5.96, respectively, using the wavelet-based algorithm, against 9.08 and 9.80 with the RSF method.

Finally, figures 4(c) and 5(c) clearly demonstrate the visual image quality improvement achieved in both oncology and brain clinical applications, permitted by the generation of PVE-corrected images using the developed algorithm. Aside from these visual improvements, ROI analyses were performed to get a quantitative insight of the correction in clinical images. In the whole-body PET image and before PVE correction, the average intensities in $ROI_{lung}$, $ROI_{heart}$ and $ROI_{lesion}$ were 35.1, 245.2 and 109.2, respectively. After correction with the proposed multiresolution method, these mean intensities changed to 36.4 (+3.7%), 240.0 (−2.1%) and 129.2 (+18.3%), respectively. This led to an increase of lesion-to-lung ratio of 16.1%. In the brain PET image, the mean intensity in white matter was 113.8 before correction and 90.7 after PVE correction. In the grey matter, the values were 126.6 before correction and 162.3 after correction, representing a 28.2% increase.

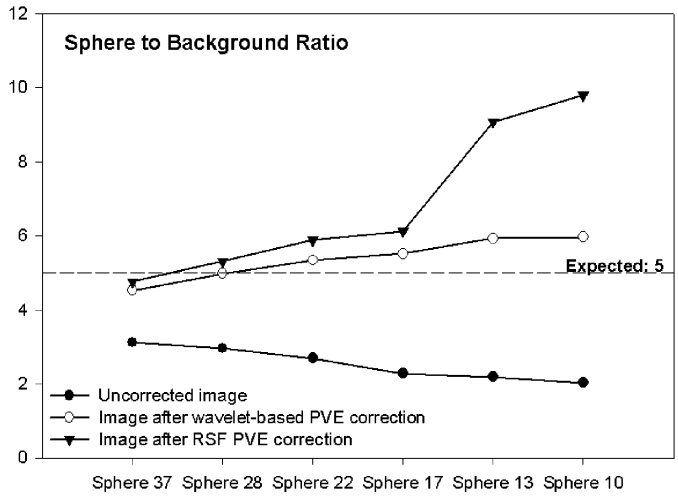



**Figure 12.** Sphere-to-background ratio in each of the different diameter spheres in the transaxial reconstructed slice of the simulated IEC phantom, before PVE correction and after PVE correction by either the RSF method or the developed wavelet-based algorithm. Expected ratios are equal to 5.

## 4. Discussion

The development of spatial co-registration algorithms in the last decade has allowed the automated and reliable superimposition of multimodality images in a day-to-day practice, in as different domains as neurology or cardiology. Moreover, since the advent of combined PET/CT and more recently SPECT/CT scanners, there has been a widespread acceptance of this new technology. This development has lead to easier and direct superimposition of functional and anatomical images (Vogel *et al* 2004) for oncology applications. Consequently, the use of anatomical data has naturally become one of the keys in addressing the problem of partial volume effects in emission tomography. However, the current PVE correction methods focus mainly on cerebral imaging, while PVE remains a major problem in other applications notably in oncology and whole-body studies. Actually, the size of tumours is often close to the PSF of PET scanners and consequently they are significantly exposed to PVE (Soret *et al* 2001). In this case, the activity and the dimension of tumours, which are critical parameters in quantitative accuracy for applications such as response to therapy or radiotherapy treatment planning (Caldwell *et al* 2003), become difficult to assess.

As previously stated, most of the different methods of correction that have been published till now suffer from the need to perform a segmentation of the anatomical information of interest and their subsequent specificity to brain imaging, with very few procedures having been tested on other organs. In addition, the vast majority of developed algorithms focus on the recuperation of accurate activity concentrations and not on yielding PVE-corrected images. The methodology developed by Rousset *et al* (1998) and referred to throughout this paper as the RSF method remains indeed the most widely used approach because it is straightforward to implement and it only requires the knowledge of the PSF. Like many other approaches it however also needs the segmentation of anatomical structures, which may become tedious and time consuming when no automatic algorithm is available. It is important as well to underline that this algorithm needs the ROIs to be accurately delineated (Frouin *et al* 2002, Zaidi *et al*

2005) which may be uncertain in the case of manual segmentation such as may be the case in whole-body PET.

In this paper, a novel approach aiming at overcoming these limitations was proposed. The general concept of the technique is a mutual multiresolution analysis of functional and anatomical images that are supposed to be correctly co-registered. During the process, the details in the high-resolution image are automatically extracted, altered according to a model and introduced in the PET image by using a simple pixel-to-pixel addition. The correction image containing these modified details is built from wavelet coefficients and consequently has a zero mean. As a result, the algorithm adjusts the intensity in boundaries of the organs and greatly enhances the smallest parts like lesions or tumours, but at the same time it hardly modifies the intensity in large and homogeneous structures.

The performance of the developed algorithm was assessed through the use of synthetic and simulated images in direct comparison with the most popular PVE correction methodology developed by Rousset *et al* in the first instance for brain imaging applications. The data obtained from the use of synthetic images demonstrated the correction ability of both aspects of PVE, that is to say tissue-fraction effects and spillover, with as good accuracy as the RSF method. In addition to the quantitatively accurate results, the developed methodology allowed the derivation of enhanced images. One may argue that the synthetic high-resolution image H had ideal properties since the L images were derived from it. Actually, the intensity in the different parts of H was rigorously the same as that we wanted to retrieve in L. However, the simple model that we defined between the wavelet coefficients of L and H allowed us to obtain similar corrected values in L by flipping the contrast in H, illustrating the robustness of the model against the nature of the high-resolution image. Similarly, good results were obtained using simulated PET images of the IEC phantom, considering the presence of lesions from 10 mm to 40 mm in diameter, with the developed methodology leading to a larger improvement in comparison to the RSF method, particularly for smaller lesions. Finally, the improvement that can be derived from the developed technique in the clinical setting was demonstrated by the enhanced PVE-corrected PET images produced as a result of the developed algorithm considering both brain and oncology applications. Such enhanced images may allow a more accurate lesion delineation providing solutions in clinical PET applications of increasing importance such as response to therapy studies and use for radiotherapy treatment planning. In the whole-body PET image, indeed, the improved visual delineation of the lesion came with an increase of intensity leading to a 16% increase of lesion-to-lung ratio. As far as brain PET is considered, the PVE correction induced compensation for both spill-in and spillover effects. The intensity in the white matter decreased in the benefit of grey matter acting as a redistribution of activity from an overestimated area to an underestimated one. The accuracy and the precision of these quantitative results are difficult to evaluate but the global behaviour of the correction is in accordance with an expected 'inverse' cross-contamination.

However, the improvement in grey matter uptake seems to be lower than that obtained in simulated images as described by Quarantelli *et al* (2004). Even if the comparison of results from real and simulated images is questionable, this point justifies some discussion. First, the proposed PVE correction operates in 2D since the wavelet transform that we use (the 'à trous' algorithm) is restricted to 2D images. This is a notable weakness but a potential solution would consist in taking into account the 3D nature of PET images, for example by performing the 'à trous' algorithm in coronal and sagittal slices in addition to the transverse plane. In the same way, the 16% increase of lesion-to-lung ratio in the whole-body PET image may appear moderate. Potential improvement in this domain could be foreseen by performing the multiresolution analysis in a limited ROI surrounding the lesion instead of the whole image containing very different tissues and organs. The simple and global linear

model that was presented in this study (allowing us to build lacking details of the PET image from the details of the CT or MRI images, see section 2.2) may indeed not be adapted to very extended regions presenting with heterogeneous structures, typically like abdominal images where clearly certain structures such as intestines appear differently in CT and PET. Therefore, although a simple and global model may be better adapted to a visual improvement alone, a quantitative study under clinical imaging conditions could potentially benefit from considering a more restricted field of view. In support of this statement, it is worth considering the results that were obtained in the second unfavourable configuration of synthetic images (figures 2(c) and (d)). In this latter case, an intense local artefact was observed after applying our algorithm, thus leading to a corrupted correction. However, if we operate our method in a limited area surrounding the given disc as demonstrated in figure 2(e), the algorithm results in the expected partial volume correction for the specified region of interest (figure 2(f)). Obviously, the remainder of the image is not PVE corrected. It is also important to note that the same kind of artefact as that shown in figure 2(d) will appear in a corrected PET image, considering the use of the global linear model, when a given tissue is present in the CT and not in the PET. A potential solution to this particular case will be the use of a restricted field of view for the application of the global model in combination with the introduction of a measure of similarity between the wavelet components of the high- and low-resolution images within the particular ROI. However, one must keep in mind that the aim is to correct for PVE in a part of the PET image corresponding to an actual tissue of interest. If there is no structure of interest present in PET, there is no associated interest in PVE correction.

Finally, only circular and homogeneous lesions were tested in the synthetic and simulated images that were presented in this paper. In clinical practice concerning whole-body imaging for example, lesions can obviously be of different shape and have non-uniform activity concentration. Here again, a more sophisticated model rather than the linear model used in the implemented algorithm in combination with its application in a more limited region may be more appropriate.

## 5. Conclusion

In this paper, a novel technique to correct emission tomography for partial volume effects using multiresolution analysis has been presented. The advantages of this approach are three-fold and can be summarized as follows. According to various tests on synthetic images, the efficiency of the correction proved to be as good as a reference method based on regional spread functions. On the other hand, the wavelet-based correction leads to better results in simulated images. In addition, and contrary to the RSF method, the process allows enhancing the images themselves to perform further processing, without any time-consuming step of ROI delineation. Finally, images of any kind of tissues, organs, functions and metabolisms are likely to be corrected, provided an anatomical image of the same object is available and correctly aligned. A simple linear and global link between the wavelet coefficients of the emission image and those of the anatomical one has been defined in this study. Under certain imaging conditions, a local model, defining only a limited area around the tissue of interest, may be more appropriate and will be considered in future developments.

## Acknowledgment

This work has been financially supported by the Region of Brittany.

## Appendix

$$5 \times 5 \text{ bicubic spline}: \begin{bmatrix} 1/256 & 1/64 & 3/128 & 1/64 & 1/256 \\ 1/64 & 1/16 & 3/32 & 1/16 & 1/64 \\ 3/128 & 3/32 & 9/64 & 3/32 & 3/128 \\ 1/64 & 1/16 & 3/32 & 1/16 & 1/64 \\ 1/256 & 1/64 & 3/128 & 1/64 & 1/256 \end{bmatrix}.$$

$$5 \times 5 \text{ linear interpolation}: \begin{bmatrix} 1/100 & 1/50 & 3/25 & 1/50 & 1/100 \\ 1/50 & 1/25 & 2/25 & 1/25 & 1/50 \\ 1/25 & 2/25 & 4/25 & 2/25 & 1/25 \\ 1/50 & 1/25 & 2/25 & 1/25 & 1/50 \\ 1/100 & 1/50 & 1/25 & 1/25 & 1/100 \end{bmatrix}.$$

$$3 \times 3 \text{ linear interpolation}: \begin{bmatrix} 1/16 & 1/8 & 1/16 \\ 1/8 & 1/4 & 1/8 \\ 1/16 & 1/8 & 1/16 \end{bmatrix}.$$

$$3 \times 3 \text{ low scale}: \begin{bmatrix} 1/172 & 1/86 & 1/172 \\ 1/86 & 160/172 & 1/86 \\ 1/172 & 1/86 & 1/172 \end{bmatrix}.$$

## References


Ardekani B A, Braun M, Hutton B F, Kanno I and Iida H 1996 Minimum cross-entropy reconstruction of PET images using prior anatomical information *Phys. Med. Biol.* **41** 2497–517

Ashburner J and Friston K J 2000 Voxel-based morphometry—the methods *Neuroimage* **11** 805–21

Aston J A, Cunningham V J, Asselin M C, Hammers A, Evans A C and Gunn R N 2002 Positron emission tomography partial volume correction: estimation and algorithms *J. Cereb. Blood Flow Metab.* **22** 1019–34

Baete K *et al* 2004 Evaluation of anatomy based reconstruction for partial volume correction in brain FDG-PET *Neuroimage* **23** 305–17

Bencherif B, Stumpf M J, Links J M and Frost J J 2004 Application of MRI-based partial-volume correction to the analysis of PET images of mu-opioid receptors using statistical parametric mapping *J. Nucl. Med.* **45** 402–8

Boussion N, Cinotti L, Barra V, Ryvlin P and Mauguiere F 2003 Extraction of epileptogenic foci from PET and SPECT images by fuzzy modeling and data fusion *Neuroimage* **19** 645–54

Caldwell C B, Mah K, Skinner M and Danjoux C E 2003 Can PET provide the 3D extent of tumor motion for individualized internal target volumes? A phantom study of the limitations of CT and the promise of PET *Int. J. Radiat. Oncol. Biol. Phys.* **55** 1381–93

Du Y, Tsui B M W and Frey E C 2005 Partial volume effect compensation for quantitative brain SPECT imaging *IEEE Trans. Med. Imaging* **24** 969–76

Dutilleux P 1987 An implementation of the "algorithme à trous" to compute the wavelet transform *Congrès ondelettes et méthodes temps-fréquence et espace des phases (Marseille, France, 14–18 September)* pp 298–304

Feuardent J, Soret M, de Dreuille O, Foehrenbach H and Buvat I 2003 Reliability of SUV estimates in FDG PET as a function of acquisition and processing protocols *IEEE Nuclear Science Symposium Conference Record* pp 2877–81

Frouin V, Comtat C, Reilhac A and Gregoire M C 2002 Correction of partial-volume effect for PET striatal imaging: fast implementation and study of robustness *J. Nucl. Med.* **43** 1715–26

Hickey K T *et al* 2004 Assessment of cardiac wall motion and ejection fraction with gated PET using N-13 ammonia *Clin. Nucl. Med.* **29** 243–8

Holdschneider R, Kronland-Martinet R, Morlet J and Tchamitchian P 1989 A real time algorithm for signal analysis with the help of the wavelet transform *Wavelets* (Berlin: Springer)

Hutton B F and Lau Y H 1998 Application of distance-dependent resolution compensation and post-reconstruction filtering for myocardial SPECT *Phys. Med. Biol.* **43** 1679–93

IEC Publication 61675-1 1998 *Radionuclide Imaging Devices—Characteristics and Test Conditions: Part 1. Positron Emission Tomographs*

Kennedy J A, Azhari H, Frenkel A, Bar-Shalom R and Israel O 2005 PET/CT image data fusion in hybrid scanners *Abstract in Proc. 52nd SNM Annual Meeting (Toronto, 18–22 June)* p 167

Kusano M L, Caldwell C B, Lanctot K L and Chow T W 2005 Evaluation and refinement of an MR-based brain PET partial volume correction method using the Zubal brain phantom *Abstract in Proc. 52nd SNM Annual Meeting (Toronto, 18–22 June)* p 452

Lamare F, Turzo A, Bizais Y, Cheze Le Rest C and Visvikis D 2006 Validation of a Monte Carlo simulation of the Philips Allegro/Gemini PET systems using GATE *Phys. Med. Biol.* at press **See endnote 1**

Li S and Shawe-Taylor J 2005 Comparison and fusion of multiresolution features for texture classification *Pattern Recognit. Lett.* **26** 633–8

Luo R C, Yih C C and Su K L 2002 Multisensor fusion and integration: approaches, applications, and future research directions *IEEE Sensors J.* **2** 107–19

Mallat S 1989 A theory for multiresolution signal decomposition: the wavelet representation *IEEE Trans. Pattern Anal. Mach. Intell.* **11** 674–93

Matsuda H, Ohnishi T, Asada T, Li Z J, Kanetaka H, Imabayashi E, Tanaka F and Nakano S 2003 Correction for partial-volume effects on brain perfusion SPECT in healthy men *J. Nucl. Med.* **44** 1243–52

Meltzer C C, Kinahan P E, Greer P J, Nichols T E, Comtat C, Cantwell M N, Lin M P and Price J C 1999 Comparative evaluation of MR-based partial-volume correction schemes for PET *J. Nucl. Med.* **40** 2053–65

Meltzer C C, Leal J P, Mayberg H S, Wagner H N Jr and Frost J J 1990 Correction of PET data for partial volume effects in human cerebral cortex by MR imaging *J. Comput. Assist. Tomogr.* **14** 561–70

Muller-Gartner H W, Links J M, Prince J L, Bryan R N, McVeigh E, Leal J P, Davatzikos C and Frost J J 1992 Measurement of radiotracer concentration in brain gray matter using positron emission tomography: MRI-based correction for partial volume effects *J. Cereb. Blood Flow Metab.* **12** 571–83

Pretorius P H and King M A 2004 Spillover and partial volume effect compensation to enhance the diagnostic accuracy of cardiac SPECT perfusion imaging *Abstract in Proc. 51st SNM Annual Meeting (Philadelphia, 18–23 June)* p 399

Quarantelli M *et al* 2004 Integrated software for the analysis of brain PET/SPECT studies with partial-volume-effect correction *J. Nucl. Med.* **45** 192–201

Ranchin T and Wald L 2000 Fusion of high spatial and spectral resolution images: the ARSIS concept and its implementation *Photogramm. Eng. Remote Sens.* **66** 49–61

Reader A J *et al* 2002 One-pass list-mode EM algorithm for high-resolution 3D PET image reconstruction into large arrays *IEEE Trans. Nucl. Sci.* **49** 693–9

Rota Kops E and Krause B J 2005 The influence of filtered back-projection and iterative reconstruction on partial volume correction in PET *Nuklearmedizin* **44** 99–106

Rousset O G, Deep P, Kuwabara H, Evans A C, Gjedde A H and Cumming P 2000 Effect of partial volume correction on estimates of the influx and cerebral metabolism of 6-[(18)F]fluoro-L-dopa studied with PET in normal control and Parkinson's disease subjects *Synapse* **37** 81–9

Rousset O G, Ma Y and Evans A C 1998 Correction for partial volume effects in PET: principle and validation *J. Nucl. Med.* **39** 904–11

Som S, Hutton B F and Braun M 1998 Properties of minimum cross-entropy reconstruction of emission tomography with anatomically based prior *IEEE Trans. Nucl. Sci.* **45** 3014–21

Somayajula S, Asma E and Leahy R 2005 PET image reconstruction using anatomical information through mutual information based priors *IEEE Nuclear Science Symposium Conference Record* pp 2722–6

Soret M, Riddell C, Hapdey S and Buvat I 2001 Biases affecting tumor uptake measurements in FDG-PET *IEEE Nuclear Science Symposium Conference Record* pp 2119–23

Starck J L, Murtagh F and Bijaoui A 1998 *Image Processing and Data Analysis: The Multiscale Approach* (Cambridge: Cambridge University Press)

Vogel W V, Oyen W J, Barentsz J O, Kaanders J H and Corstens F H 2004 PET/CT: panacea, redundancy, or something in between? *J. Nucl. Med.* Suppl. 1 **45** 15S–24S

Wells W M III, Viola P, Atsumi H, Nakajima S and Kikinis R 1996 Multi-modal volume registration by maximization of mutual information *Med. Image Anal.* **1** 35–51

Wen C Y and Chen J K 2004 Multi-resolution image fusion technique and its application to forensic science *Forensic Sci. Int.* **140** 217–32

Zaidi H, Ruest T, Schoenahl F and Montandon M L 2005 Comparative assessment of brain MR image segmentation algorithms and their impact on partial volume effect correction in PET *Abstract in Proc. 52nd SNM Annual Meeting (Toronto, 18–22 June)* p 454

## Endnotes

(1) Author: Please update reference 'Lamare *et al* (2006)'.

---

Available online at www.sciencedirect.com
ScienceDirect
Pattern Recognition Letters 28 (2007) 2201–2208
Pattern Recognition Letters
www.elsevier.com/locate/patrec

# Non-stationary fuzzy Markov chain

F. Salzenstein [a,*], C. Collet [b], S. Lecam [b], M. Hatt [c]

[a] Laboratoire INESS, UMR CNRS 7163, Université Strasbourg 1 (ULP), France
[b] LSIIT UMR CNRS 7005, Université Strasbourg 1 (ULP), France
[c] Laboratoire LATIM, INSERM U650, Université de Brest (UBO), France

Received 18 July 2006; received in revised form 17 April 2007
Available online 17 July 2007

Communicated by J.A. Robinson

**Abstract**

This paper deals with a recent statistical model based on fuzzy Markov random chains for image segmentation, in the context of stationary and non-stationary data. On one hand, fuzzy scheme takes into account discrete and *continuous* classes through the modeling of hidden data imprecision and on the other hand, Markovian Bayesian scheme models the uncertainty on the observed data. A non-stationary fuzzy Markov chain model is proposed in an unsupervised way, based on a recent Markov triplet approach. The method is compared with the stationary fuzzy Markovian chain model. Both stationary and non-stationary methods are enriched with a parameterized joint density, which governs the attractiveness of the neighbored states. Segmentation task is processed with Bayesian tools, such as the well known MPM (Mode of Posterior Marginals) criterion. To validate both models, we perform and compare the segmentation on synthetic images and raw optical patterns which present diffuse structures.


*Keywords:* Fuzzy Markov chain; Triplet Markov chain; Non-stationary chain; Multispectral image segmentation

## 1. Introduction

The fuzzy segmentation problem consists of estimating the hidden realization $x = (x_s)_{1 \leqslant s \leqslant N}$, for a given set of $D$ observations $\boldsymbol{Y} = \boldsymbol{y} = \{\boldsymbol{y}_s \in \mathbb{R}^D\}$, where $x_s = (\varepsilon_1(s), \varepsilon_2(s), \ldots, \varepsilon_K(s))$. Each component $\varepsilon_i(s)$ represents the contribution of each class $\omega_i$ in a finite discrete set $\Omega = \{\omega_1, \ldots, \omega_K\}$ of $K$ hard classes. The fuzzy belonging of each pixel respects the normalization condition: $\varepsilon_1(s) + \varepsilon_2(s) + \cdots + \varepsilon_K(s) = 1$. In the context of two "hard" classes, a set $\Omega = \{0, 1\}$ yields $x_s \in [0, 1]$. Then, all values $x_s \in [0, 1]$ model the proportion of the class "0" in the pixel related to $X_s$, whereas $1 - x_s$ corresponds to the proportion of the class "1". The distribution at each random variable $X_s$ is given by a density $h_s$ with respect to a measure $v$ including discrete components (Dirac functions $\delta_0$, $\delta_1$ on $\{0, 1\}$) and a continuous component (the Lebesgue measure $\mu$ on $]0, 1[$) (Caillol et al., 1993):

$$v = \delta_0 + \delta_1 + \mu. \tag{1}$$

The discrete components of $v$ are associated with the hard classes, whereas the continuous component $\mu$ is associated with the fuzzy feature. In this paper, we will consider the case $D = N$ (mono-spectral context). When $X$ is a Markov chain called "fuzzy Markov chain" (FMC) and the variable $\boldsymbol{Y}$ is independent conditionally on $X$, it is possible to express the joint distribution $p(x, \boldsymbol{y})$ with respect to a measure $v^N \otimes \mu^N$, as follows:

$$p(x, \boldsymbol{y}) = p(x_1)p(x_2|x_1) \cdots p(x_N|x_{N-1})p(\boldsymbol{y}_1|x_1) \cdots p(\boldsymbol{y}_N|x_N). \tag{2}$$

In particular, the *posterior* field $X$ conditional on $\boldsymbol{Y}$ is Markovian. Thus, one can process the posterior realizations of

* Corresponding author. Tel.: +33 3 88 10 65 58; fax: +33 3 88 10 65 48.
*E-mail addresses:* Fabien.Salzenstein@iness.c-strasbourg.fr (F. Salzenstein), Christophe.Collet@lsiit.u-strasbg.fr (C. Collet), Steven.Lecam@lsiit.u-strasbg.fr (S. Lecam), Mathieu.Hatt@univ-brest.fr (M. Hatt).

0167-8655/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2007.07.002

the hidden variable $X$, called Hidden Fuzzy Markov Chain (HFMC). Generally the distribution $p(x, \boldsymbol{y})$ depends on unknown parameter $\theta = (\theta_X, \theta_Y)$ where the *prior* parameters $\theta_X$ define the prior density of the Markov chain and the parameters $\theta_Y$ define the distribution parameters of the driven data conditional on $X$. Algorithms like "Expectation Maximization" (EM) (McLachlan and Krishnan, 1997) or its stochastic version (SEM) (Celeux and Diebolt, 1985) are efficient to estimate the hyper-parameter when $\theta_X$ does not vary locally, i.e., when the variable $X$ is stationary. Recent studies have focused on unsupervised segmentation of Markov chain in the fuzzy context (Avrachenkov and Sanchez, 2002; Mohammed and Gader, 2000; Carincotte et al., 2004). In particular, we derive $\theta_X$ from the prior joint density $p(x_s, x_{s+1})$ at each neighbored sites. We present here a new model based on a parameterized joint density, which governs locally the attractiveness between two neighbored states. Unfortunately, when $\theta_X$ does not vary locally, these approaches are sometimes badly adapted. Thus one has to introduce a new fuzzy hidden Markov chain model which represents non-stationary data. In this work we model the non-stationarity by a third auxiliary process $U$, which governs the changing values of $\theta_X$ in the hidden process. A such method has been successfully applied in the hard context (Hughes et al., 1999; Lanchantin and Pieczynski, 2004), and we propose in this article to extend non-stationary Markov chain to the fuzzy case. The solution proposed in (Lanchantin and Pieczynski, 2004) is derived from a recent triplet Markov chain model (Pieczynski, 2002) which can be described in the following manner: the pairwise process $Z = (X, U)$ is assumed to be Markovian, $X$ and $U$ separately are not necessary Markovian. The triplet process $T = (X, U, Y)$ is then a particular triplet Markov chain. In Section 2, we present the stationary fuzzy Markov chain (SFMC) with and without a parameterized joint density (P-SFMC versus NP-SFMC). We briefly introduce the stationary fuzzy Markov field (SFMF) (Salzenstein and Pieczynski, 1997), which is used in the experimental part to enrich the comparisons. In the next Section3 we generalize the non-stationary model of Lanchantin and Pieczynski (2004) presenting a new fuzzy model in the context of non-stationary Markov chain (NSFMC) with a possibly joint parameterized density (P-NSFMC versus NP-NSFMC). We describe the noise model used (Section 4), the MPM segmentation procedure applied to the S/NS-FMC methods (Section 5) and the associated hyper-parameter estimation step (Section 6). Finally we show the efficiency of the new method though synthetic images (Section 7) and real images (Section 8).

## 2. The stationary fuzzy Markov chain (SFMC)

Let us consider now a Markov chain $X = (x_s)_{1 \leqslant s \leqslant N}$ with continuous statements, i.e., $X_s \in [0, 1]$. To define the distribution $\pi(x)$ of the variable $X$, we need the density $p(x_1)$ of the initial distribution, and the transition densities $p(x_s|x_{s-1})_{1 \leqslant s \leqslant N}$:

$$p(x_1, x_2, \ldots, x_N) = p(x_1) \cdot p(x_2|x_1) \cdots p(x_N|x_{N-1}). \tag{3}$$

When the chain is stationary, all prior distributions can be deduced from a joint density. The prior joint density $p(x_s, x_{s+1})$ is defined on the pairwise $(x_s, x_{s+1}) \in [0, 1]^2$. According to a measure $v \otimes v$, the normalization condition yields

$$\int_0^1 \int_0^1 p(u, v) d(v \otimes v)(u, v) = 1. \tag{4}$$

We propose a general model to define it

$$\begin{aligned} p(\varepsilon_1, \varepsilon_2) &= a \cdot \phi(\varepsilon_1, \varepsilon_2) + b \quad \text{with } \phi(\varepsilon_1, \varepsilon_2) \\ &= \phi(\varepsilon_2, \varepsilon_1); \quad (a, b) \in \mathbb{R}^2. \end{aligned} \tag{5}$$

The function $\phi(\varepsilon_1, \varepsilon_2)$, is applied when at least one label is fuzzy, i.e., $\varepsilon_1$ or $\varepsilon_1 \in ]0, 1[$. If both labels are hard, we note $p(0, 0) = \pi_{00},\ p(1, 1) = \pi_{11},\ p(0, 1) = \pi_{01},\ p(1, 0) = \pi_{10}$. We model a parameterized function $\phi$ as follows:

$$\phi(\varepsilon_1, \varepsilon_2) = (1 - |\varepsilon_1 - \varepsilon_2|)^r \quad r \in \mathbb{R}. \tag{6}$$

When $r$ increases, the probability of having two similar neighbored pixels increases: thus, the parameter $r$ governs the homogeneity of the image, i.e., the attractiveness between the different states. Moreover, the limit conditions that we impose yield

$$p(0, 1) = p(1, 0) = b = \pi_{01} = \pi_{10}. \tag{7}$$

Applying Eq. (4) yields to the general condition:

$$\begin{aligned} &p(0, 0) + p(1, 1) + p(0, 1) + p(1, 0) + 2 \cdot \int_{]0,1[} p(0, u) \mathrm{d}u \\ &\quad + 2 \cdot \int_{]0,1[} p(1, u) \mathrm{d}u + \int_{]0,1[} \int_{]0,1[} p(u, v) \mathrm{d}u \, \mathrm{d}v = 1. \end{aligned} \tag{8}$$

This gives a relationship between all prior parameters $(\pi_{00}, \pi_{11}, \pi_{01}, \pi_{10}, a, b)$. When the prior joint density is defined by the function (6), we compute (8) using a quantization of the interval $[0, 1]$ into $M$ equidistant values: $\{\varepsilon_0 = 0,\ \varepsilon_1 = \frac{1}{M} \ldots,\ \varepsilon_i = \frac{i}{M}, \ldots,\ \varepsilon_M = 1\}$. Then, we derive the initial density $p(x_1)$, which corresponds to the marginal distribution (9)

$$\begin{aligned} p(x_s) &= \int_0^1 p(x_s, \varepsilon) \mathrm{d}v(\varepsilon) \\ &= p(x_s, 0) + p(x_s, 1) + \int_{]0,1[} p(x_s, \varepsilon) \mathrm{d}\varepsilon. \end{aligned} \tag{9}$$

In this section we briefly described a new fuzzy Markov random chain model based on a parameterized joint density associated to the transition probabilities. The stationary fuzzy chain associated with a non-parameterized density is named NP-SFMC. The stationary fuzzy chain associated with a parameterized density is named P-SFMC. It is also possible to define in the same way, the stationary Markovian random field (SFMF) (Salzenstein and Pieczynski, 1997) $X \in [0, 1]^N$, for which the distribution $p_X$ with respect to a measure $v^N$ is given by Salzenstein and Pieczynski (1997)

$$p_X(x) = \frac{1}{Z} \cdot \mathrm{e}^{-U_{\mathrm{f}}(x)}, \tag{10}$$

where the fuzzy energy $U_{\mathrm{f}}$ is a sum of functions $\Phi_C$ defined on neighbored sites:

$$U_{\mathrm{f}}(x) = \sum_{(x_s,x_t)\in C} \Phi_C(x_s, x_t), \tag{11}$$

where $(x_s, x_t) \in [0,1]^2$ represents a pairwise of neighbored pixels (Geman and Geman, 1984) in the image (vertical, diagonal or horizontal neighborhood), and the associated functions $\Phi_C$ defined on $[0,1]^2$. Let us notice, the SFMF procedure is time consuming: on has to compute several realizations of $X$ according to its posterior distribution, using a Gibbs sampling (Geman and Geman, 1984). In the opposite way, the Markov chain model provides an efficient tool: one can compute directly the posterior density using the forward/backward procedure (Devijver, 1985). In order to use the HMC procedure, it is possible to extract the 2D signal as mono-dimensional data using the Hilbert-Peano path. This technique preserves the neighborhood information. Let us describe now a non-stationary fuzzy Markov chain model.

## 3. The non-stationary fuzzy Markov chain (NSFMC)

Authors (Lanchantin and Pieczynski, 2004) propose to add to an initial process $X$ an additional process $U$, which takes its values in a finite set $\Lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_K\}$. The couple $Z = (X, U) = \{(x_1, u_1), (x_2, u_2), \ldots, (x_N, u_N)\}$ is supposed to be a stationary Markov chain, where $X$ is an interested non-stationary process, and $U$ models auxiliary states:

$$p(z_s = (x_s, u_s)|z_{s-1}, z_{s-2}, \ldots, z_1) = p(z_s|z_{s-1}). \tag{12}$$

In (Lanchantin and Pieczynski, 2004), $X$ and $U$ take their values into discrete classes. We propose to generalize this model by labeling each component $X_s$ into a continuous set $[0,1]$. The intermediate variable $U$ takes its values into a finite set, in order to define stationary partitions of the variable $X$. The chain $Z$ is defined by a prior joint density

$$p(z_s, z_{s+1}) = p(x_s, x_{s+1}, u_s, u_{s+1}),$$

according to the measure $\left(v + \sum_{n=1}^{K} \delta_{\lambda_n}\right) \otimes \left(v + \sum_{n=1}^{K} \delta_{\lambda_n}\right)$, the initial probability is computed by

$$\begin{aligned} &P(X_s \in I_s, X_{s+1} \in I_{s+1}, U_s = \lambda_s, U_{s+1} = \lambda_{s+1}) \\ &\quad = \int_{I_s} \int_{I_{s+1}} p(\epsilon, \eta, \lambda_s, \lambda_{s+1}) d(v \otimes v)(\epsilon, \eta), \end{aligned} \tag{13}$$

with $I_s \subset [0,1]$ and $I_{s+1} \subset [0,1]$. As in the stationary case, it is possible to define a parameterized and non-parameterized joint density, provided that the normalization condition (14) is established:

$$\sum_{\lambda_i} \sum_{\lambda_j} \int_0^1 \int_0^1 p(\epsilon, \eta, \lambda_i, \lambda_j) d(v \otimes v)(\epsilon, \eta) = 1. \tag{14}$$

In particular, this condition is written

$$\sum_{\lambda_i} \sum_{\lambda_j} P_{ij} = 1, \tag{15}$$

where

$$\int_0^1 \int_0^1 g(\epsilon, \eta, \lambda_i, \lambda_j) d(v \otimes v)(\epsilon, \eta) = P[\lambda_i, \lambda_j] = P_{ij}. \tag{16}$$

Thus, it is possible to construct a parameterized mode of $p(z_s, z_{s+1})$ by the means of the parameters $\pi_{00}^{ij}$, $\pi_{01}^{ij}$, $\pi_{10}^{ij}$, $\pi_{11}^{ij}$

$$p(0,0,\lambda_i,\lambda_j) = \pi_{00}^{ij}, \quad p(0,1,\lambda_i,\lambda_j) = \pi_{01}^{ij}, \tag{17}$$

$$p(1,0,\lambda_i,\lambda_j) = \pi_{10}^{ij}, \quad p(1,1,\lambda_i,\lambda_j) = \pi_{11}^{ij}. \tag{18}$$

When $\epsilon_1$ or $\epsilon_2 \in ]0,1[$, let us express $a_{ij}$, $b_{ij}$ and the auxiliary function $\phi$ defined by (6)

$$p(\epsilon_1, \epsilon_2, \lambda_i, \lambda_j) = a_{ij} \cdot \phi(\epsilon_1, \epsilon_2) + b_{ij}. \tag{19}$$

Moreover, we impose limit conditions (20)

$$a_{ij} \cdot \phi(0,1) + b_{ij} = a_{ij} \cdot \phi(1,0) + b_{ij} = b_{ij} = \pi_{01}^{ij} = \pi_{10}^{ij}. \tag{20}$$

Finally the neighborhood prior density depends on $4 \times K^2$ parameters $\pi_{00}^{ij}$, $\pi_{01}^{ij}$, $\pi_{11}^{ij}$. The other parameters $\pi_{10}^{ij}$, $b_{ij}$ and $a_{ij}$ are computed by the conditions (15), (16), (20). The non-stationary Markov chain based on a parameterized joint density will be named P-NSFMC whereas the model based on a non-parameterized density is named NP-NSFMC. In this section we introduced a new fuzzy non-stationary Markov random chain model. We defined the associated prior joint density, initial and transition probabilities. Let us now describe the segmentation task.

## 4. Model of the observations in a non-stationary context

The joint process $Z = (X, U)$ being assumed to be Markovian, the aim of our paper is to process multispectral data. We observe $D$ realizations $(y^{(1)}, y^{(2)}, \ldots, y^{(D)})$ of the random vector $\boldsymbol{Y} = (Y^{(1)}, Y^{(2)}, \ldots, Y^{(D)})$. They represent a single scene observed at different wavelengths or from different sensors. For each field $\boldsymbol{Y^{(i)}}$, the variables $\boldsymbol{Y^{(i)}} = \{Y_1^{(i)}, Y_2^{(i)}, \ldots, Y_N^{(i)}\}$ are spatially independent conditionally on $Z$. One has the following relationships (Lanchantin and Pieczynski, 2004):

$$p(\boldsymbol{Y^{(i)}}|Z) = \prod_{s=1}^{N} p(\boldsymbol{Y_s^{(i)}}|Z), \tag{21}$$

$$p(\boldsymbol{Y_s^{(i)}}|Z) = p(\boldsymbol{Y_s^{(i)}}|Z_s), \tag{22}$$

$$p(\boldsymbol{Y_s^{(i)}}|Z_s) = p(\boldsymbol{Y_s^{(i)}}|X_s). \tag{23}$$

According to the third condition, the data driven parameter are stationary. Hence the parameter $\theta_X$ varies locally, whereas the parameter $\theta_Y$ stays global. The distribution $f_{x_s}(y_s)$ of $\boldsymbol{y}_s$ according to $X_s = \varepsilon_s \in [0,1]$ is a Gaussian multivariate density (Salzenstein and Collet, 2006)

$$f_{x_s}(y_s) = \frac{1}{2\pi^{D/2}(\det \Gamma_{\varepsilon_s})^{1/2}} \exp^{\left(\frac{-1}{2}(y_s - \boldsymbol{\mu}_{\varepsilon_s})^t \boldsymbol{\Gamma}_{\varepsilon_s}^{-1}(y_s - \boldsymbol{\mu}_{\varepsilon_s})\right)} \tag{24}$$

where $\boldsymbol{\mu}_{\varepsilon_s} = [\mu_{\varepsilon_s}^{(1)}, \ldots, \mu_{\varepsilon_s}^{(D)}]^{\mathrm{t}}$ and $\boldsymbol{\Gamma}_{\varepsilon_s} \in \mathbb{R}^D \times \mathbb{R}^D$, respectively, define a mean vector and variance–covariance matrix, at each fuzzy/hard site. Let be $(\boldsymbol{\mu}_0, \boldsymbol{\mu}_1)$ and $(\boldsymbol{\Gamma}_0, \boldsymbol{\Gamma}_1)$ the mean vectors and variance–covariance matrix related to the hard classes "0" and "1". For each fuzzy site $X_s = \varepsilon_s$, the related mean vector and covariance matrix $\boldsymbol{\mu}_{\varepsilon_s}$ and $\boldsymbol{\Gamma}_{\varepsilon_s}$ are written

$$\boldsymbol{\mu}_{\varepsilon_s} = (1-\varepsilon_s)\cdot\boldsymbol{\mu}_0 + \varepsilon_s\cdot\boldsymbol{\mu}_1, \tag{25}$$

$$\boldsymbol{\Gamma}_{\varepsilon_s} = (1-\varepsilon_s)^2\cdot\boldsymbol{\Gamma}_0 + \varepsilon_s^2\cdot\boldsymbol{\Gamma}_1. \tag{26}$$

## 5. Segmentation procedure

### 5.1. Segmentation of the SFMC

Given the set of observations $\boldsymbol{Y} = \boldsymbol{y}$, we wish to estimate one realization $X = x \in [0,1]^N$. It is possible to adapt the MPM criterion (Maroquin et al., 1987) to the fuzzy context (Salzenstein and Pieczynski, 1997). For a such approach, the final decision process is performed as following: given a realization $\boldsymbol{Y} = \boldsymbol{y}$, the bayesian decision $\widehat{d}_s$ such that $\widehat{d}_s(Y)) = \widehat{X_s}$, will involve minimizing a conditional expectation (27) at each location $s$, in order to obtain an optimal value of $X_s$

$$\hat{x}_s^{\mathrm{opt}} = \arg\min_{\widehat{X_s}=\hat{x}_s} E[L_s(X_s, \widehat{X_s})|Y = y]. \tag{27}$$

The loss function $L_s^*(\widehat{x_s}, x_s)$ models the severity of attributing the value $\widehat{x_s}$ instead of a true one $x_s$ to the pixel. Although there are numerous possibilities in the choice of the loss function, the 'absolute distance' $L_s(x_s, \hat{x}_s) = |x_s - \hat{x}_s|$, gives efficient results for segmentation tasks. For a stationary variable $X$, the error rate is approximated by

$$E[L_s(X_s, \widehat{X}_s)] \simeq \frac{1}{N}\sum_{s=1}^{s=N} L_s(x_s, \hat{x}_s). \tag{28}$$

The calculus of (27) requires the knowledge of the posterior distribution $p_{X_s}^Y$ at each $X_s$:

$$E[L_s(X_s, \widehat{X_s})|y] = p_{X_s}^y(0)\cdot L_s(0, \widehat{X_s}) + p_{X_s}^y(1) L_s(1, \widehat{X_s}) + \int_{]0,1[} p_{X_s}^y(t) L_s(t, \widehat{X_s})\mathrm{d}t. \tag{29}$$

Segmentation is performed by affecting to each pixel a value $\widehat{X_s} \in [0,1]$ which minimizes (29). When the hidden process $X$ is a Markov chain, one can compute the posterior density using the forward/backward procedure (Devijver, 1985) in the fuzzy context. The forward and backward densities $\alpha_s(x_s), \beta_s(x_s)$ are defined by

$$\alpha_s(x_s) = p(x_s, y_1, \ldots, y_s), \tag{30}$$

$$\beta_s(x_s) = \frac{p(y_{s+1}, \ldots, y_N|x_s)}{p(y_{s+1}, \ldots, y_N|y_1, \ldots, y_s)}. \tag{31}$$

The recurrence formula providing these quantities, are analogous to the hard segmentation processing

$$\alpha_s(x_s) \propto p(y_s|x_s)\int_0^1 \alpha_{s-1}(u)\cdot p(x_s|u)\mathrm{d}v(u), \tag{32}$$

$$\beta_s(x_s) \propto \int_0^1 \beta_{s+1}(u) p(y_{s+1}|u) p(u|x_s)\mathrm{d}v(u). \tag{33}$$

The relationship $\alpha_s(\varepsilon)\cdot\beta_s(\varepsilon) = p_{X_s}^Y$ gives immediately the minimization of (29). Moreover, as in the hard context, it is possible to simulate posterior realizations of the variable $X$ using the posterior transition $p^{y_{s+1}}(x_{s+1}|x_s)$ and initial densities $p_{X_s}^Y(x_s)$ i.e., for any $(x_s, x_{s+1}) \in [0,1]^2$:

$$p_{X_s}^Y(x_s) = \alpha_s(x_s)\cdot\beta_s(x_s), \tag{34}$$

$$p^{y_{s+1}}(x_{s+1}|x_s) = \frac{p(x_{s+1}|x_s)\cdot p(y_{s+1}|x_{s+1})\cdot\beta_{s+1}(x_{s+1})}{\int_0^1 p(x|x_s)p(y_{s+1}|x)\cdot\beta_{s+1}(x)\mathrm{d}v(x)}. \tag{35}$$

### 5.2. Segmentation of the NSFMC

The segmentation problem consists then in estimating a realization of an hidden process $X = x$, which is not necessary Markovian, given a set $\boldsymbol{Y} = \boldsymbol{y}$. In order to estimate the fuzzy process $X$, one has to minimize the conditional expectation (27). In order to estimate the hidden auxiliary process $U$, we apply the classic decision task (36) corresponding to the local "0–1" loss function:

$$\widehat{U}_s^{\mathrm{opt}} = \arg\min_{\widehat{U_s}=\hat{u}_s} P[U_s|Y = y]. \tag{36}$$

Thus, it is necessary to compute the posterior densities $p[X_s|\boldsymbol{Y} = \boldsymbol{y}]$ and $p[U_s|\boldsymbol{Y} = \boldsymbol{y}]$ in order to perform the decision processes (27) and (36). They correspond to the marginalization of the distribution $p(Z_s|\boldsymbol{y})$:

$$p(x_s|y) = \sum_{\Lambda} p(z_s|y), \qquad p(u_s|y) = \int_0^1 p(z_s|y)\mathrm{d}v, \tag{37}$$

$Z = (X, U)$ being a pairwise Markov chain, it is possible to compute the posterior distribution $p(z_s|\boldsymbol{y})$ by the means of the forward–backward procedure, extended to the fuzzy context:

$$\alpha_s(z_s) \propto p(y_s|z_s)\sum_{\lambda\in\Lambda}\int_0^1 \alpha_{s-1}(\varepsilon,\lambda)\cdot p(z_s|\varepsilon,\lambda)\mathrm{d}v(\varepsilon), \tag{38}$$

$$\beta_s(z_s) \propto \sum_{\lambda\in\Lambda}\int_0^1 \beta_{s+1}(\varepsilon,\lambda) p(y_{s+1}|\varepsilon,\lambda) p(\varepsilon,\lambda|z_s)\mathrm{d}v(\varepsilon). \tag{39}$$

Using the hypothesis related to the observed data, we simplify this procedure in the following manner:

$$\alpha_s(z_s) \propto p(y_s|x_s)\sum_{\lambda\in\Lambda}\int_0^1 \alpha_{s-1}(\varepsilon,\lambda)\cdot p(z_s|\varepsilon)\mathrm{d}v(\varepsilon), \tag{40}$$

$$\beta_s(z_s) \propto \sum_{\lambda\in\Lambda}\int_0^1 \beta_{s+1}(\varepsilon,\lambda) p(y_{s+1}|\varepsilon) p(\varepsilon,\lambda|z_s)\mathrm{d}v(\varepsilon). \tag{41}$$

At each step, we compute the posterior distribution

$$p(x_s, u_s|y) = \alpha_s(x_s, u_s)\cdot\beta_s(x_s, u_s). \tag{42}$$

In the same manner as we have seen in Section 5.1, it is possible to simulate hidden realizations of the variable $Z=(X,U)$ according to the posterior initial and transition densities. This allows us to estimate hyper-parameters by using the well known SEM procedure.

## 6. Hyper-parameter estimation

We focus in this section on the estimation of the parameter $\theta$ in the context of a non-stationary variable. Actually, the stationary context is a particular case, for which $U$ owns one discrete state i.e., $Card\,\Lambda=1$. The final segmentation step requires the parameter set $\theta=(\theta_Z,\theta_Y)$ where the *prior* parameters $\theta_Z$ define the prior density of the Markov chain $Z$, which could be the set of parameters $(\pi_{00}^{ij},\pi_{01}^{ij},\pi_{10}^{ij},\pi_{11}^{ij},a_{ij},b_{ij})$ for the P-SFMC and P-NSFMC approaches. The parameters $\theta_Y=((\mu_0,\mu_1);(\sigma_0,\sigma_1))$ define the distribution of the data driven conditional on $X$. For each posterior realization of the field $Z=(X,U)$, an SEM estimator (empirical frequencies and moments) is used to estimate the hyper-parameters. When the sequence $\theta^{[p]}$ approaches steady state – for example 1% of the relative change in the values – we stop the procedure. Let us consider now the problem in estimating $\theta_Z$ and $\theta_Y$ separately.

### 6.1. Data driven parameter estimation

Let us suppose now, we observe a realization $(x,\boldsymbol{y})$ of the pairwise $(X,\boldsymbol{Y})$. In a hard classification the empirical moment estimator $\hat{\theta}_Y(X,Y)$ of $\theta_Y$ corresponds to the maximum likelihood under conditional Gaussian laws assumption. When we use a fuzzy classification, it is enough to estimate the parameters dealing with hard classes. We generalize the method proposed in (Salzenstein and Pieczynski, 1997) applying the empirical moments to the hard pixels. Let be $Q_p=\{s\in S/X_s=p\}, p=0,1$ the sets of pixels which belong to the hard classes. Our aim is to estimate the set of parameters $\theta_Y=(\mu_0,\boldsymbol{\Gamma}_0;\mu_1,\boldsymbol{\Gamma}_1)=(\mu_0^{(i)},\Gamma_0^{(i,j)};\mu_1^{(i)},\Gamma_1^{(i,j)})$, where $1<i,j<D$. Applying the empirical moment method on the hard pixels yields

$$\widehat{\mu}_p^{(i)}=\frac{\sum_{s\in Q_p}y_s^{(i)}\cdot\delta(x_s,p)}{\sum_{s\in Q_p}\delta(x_s,p)}, \tag{43}$$

$$\widehat{\Gamma}_p^{(i,j)}=\frac{\sum_{s\in Q_p}(y_s^{(i)}-\widehat{\mu}_p^{(i)})\cdot(y_s^{(j)}-\widehat{\mu}_p^{(j)})\delta(x_s,p)}{\sum_{s\in Q_p}\delta(x_s,p)}. \tag{44}$$

### 6.2. Prior parameter estimation

Let us consider an hidden chain $Z=(X,U)$ simulated by its posterior distribution, according to the procedure described in Section 5.1. The prior parameter $\theta_Z$ corresponds to the initial and transition densities. They can be deduced from the joint density, as seen in Section 3. We consider two hypothesis: (i) the joint density is a non-parameterized density. (ii) It depends on a parameterized function $\phi$.

(i) NP-NSFMC: We compute the empirical $M^2\times K^2$ joint probabilities (45) according to different neighborhood configurations

$$P[X_s\in I_i,X_{s+1}\in I_j,U_s=\lambda_p,U_{s+1}=\lambda_q]=\int_{I_i}\int_{I_j}p(\varepsilon_s,\varepsilon_{s+1},\lambda_p,\lambda_q)d(v\otimes v)\varepsilon_s,\varepsilon_{s+1}) \tag{45}$$

We deduce the joint density from these probabilities. For instance, when $I_i=\left[\frac{i}{M},\frac{i+1}{M}\right[$ and $I_j=\left[\frac{j}{M},\frac{j+1}{M}\right[$ corresponding to a discretization of $[0,1]$

$$p\left(\epsilon_i=\frac{i}{M},\epsilon_j=\frac{j}{M},\lambda_p,\lambda_q\right)\simeq\frac{1}{M^2}\cdot P[X_s\in I_i,X_{s+1}\in I_j,U_s=\lambda_p,U_{s+1}=\lambda_q]. \tag{46}$$

(ii) P-NSFMC: when the joint density is parameterized, we have to estimate the quantities $\pi_{00}^{ij}$, $\pi_{01}^{ij}$, $\pi_{11}^{ij}$ and $P_{ij}$ according to the empirical frequencies

$$\pi_{00}^{ij}=\frac{\sum_{s=1}^{N-1}1_{[(x_s,u_s)=(0,i);(x_{s+1},u_{s+1})=(0,j)]}}{N}, \tag{47}$$

$$\pi_{01}^{ij}=\frac{\sum_{s=1}^{N-1}1_{[(x_s,u_s)=(0,i);(x_{s+1},u_{s+1})=(1,j)]}}{N}, \tag{48}$$

$$\pi_{11}^{ij}=\frac{\sum_{s=1}^{N-1}1_{[(x_s,u_s)=(1,i);(x_{s+1},u_{s+1})=(1,j)]}}{N}, \tag{49}$$

$$P_{ij}=\frac{\sum_{s=1}^{N-1}1_{[u_s=i;u_{s+1}=j]}}{N}. \tag{50}$$

$\pi_{01}^{ij}$, $a_{ij}$ and $b_{ij}$ are deduced by conditions (15), (16), (20).

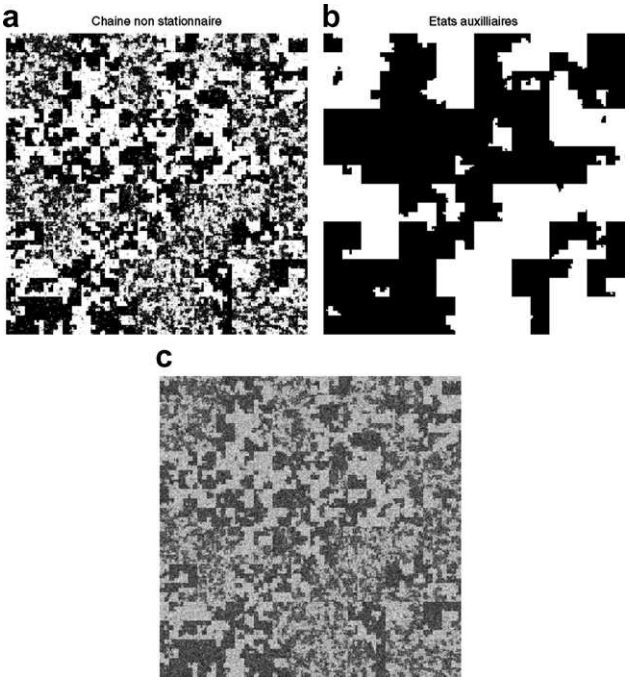



Fig. 1. (a) A non-stationary fuzzy Markov chain $X$; (b) its related states $U$; (c) its noisy version $Y$.

## 7. Results on synthetic images

We simulated a non-stationary fuzzy Markov chain on $M = 10$ discrete fuzzy levels, with two homogeneous states ($Card\,\Lambda = 2$) and $r = 1$. The variables $X$ and $U$ are represented in Fig. 1a and b. The class "0" (in black) of $U$ corresponds to an hard-dominating area in $X$ ($\pi_{00}^{11} = \pi_{11}^{11} = 0.2$), where as the class "1" (in white) corresponds to a fuzzy area in $X$ ($\pi_{00}^{11} = \pi_{11}^{11} = 0.05$). A noisy version is represented in Fig. 1c. We give below the following corresponding prior and data driven parameters:

$$\pi_{00}^{ij} = \pi_{11}^{ij} = \begin{pmatrix} 0.2 & 0 \\ 0 & 0.05 \end{pmatrix} \quad P_{ij} = \begin{pmatrix} 0.4995 & 0.005 \\ 0.005 & 0.4995 \end{pmatrix}$$

$$\pi_{10}^{ij} = \pi_{01}^{ij} = b_{ij} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$$

$$(\mu_0, \mu_1) = (120, 142) \quad (\sigma_0, \sigma_1) = (4, 4).$$

Fig. 2a and b gives the segmented fields $X$ corresponding to the non-stationary ($Card\,\Lambda = 2$) case when one considers respectively the parameterized (P-NSFMC with $r = 1$) and non-parameterized (NP-NSFMC) approaches. Fig. 2c–e corresponds to the stationary algorithm ($Card\,\Lambda = 1$) applied to the image, respectively by the means of a parameterized (P-SFMC), non-parameterized (NP-SFMC) fuzzy Markov chain. Another stationary method used is based on a fuzzy Markov field (SFMF), briefly presented in Section 2. For the non-stationary approaches with two states, the segmented field $U$ is represented in Fig. 3a and b. Moreover, the estimated prior parameters are indicated in Table 1, where as the noise parameters for NSFMC, SFMC and SFMF are given, respectively, in Tables 2 and 3. Finally the error rates computed by (28) are given for all NP/P-(N)SFMC and SFMF procedures in Table 4. The stationary method based on FMC or FMF give higher rates of error than the non-stationary methods. The highest rate is given by the SFMF.

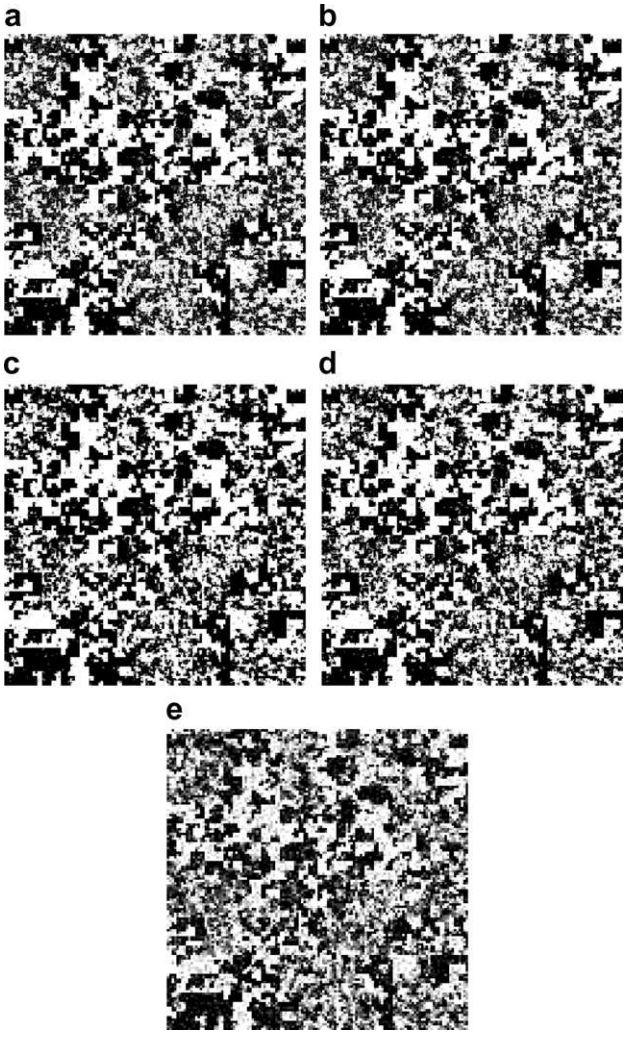



Fig. 2. Segmented images of $X$. (a) P-NSFMC; (b) NP-NSFMC; (c) P-SFMC; (d) NP-SFMC; (e) SFMF.

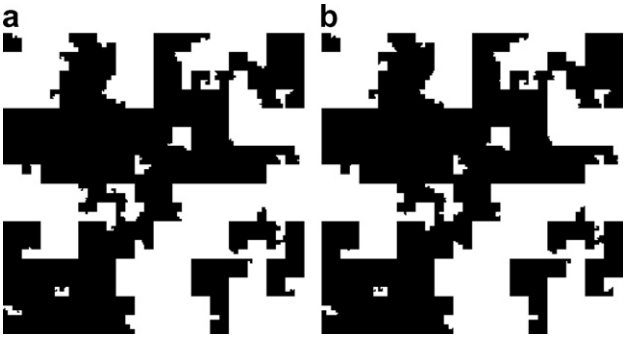



Fig. 3. Segmented images of $U$. (a) P-NSFMC; (b) NP-NSFMC.

Table 1
Estimated prior parameters for the non-stationary procedures

| Parameters | P-NSFMC | NP-NSFMC |
|---|---|---|
| $\hat{\pi}_{00}$ | $\begin{pmatrix} 0.23 & 0 \\ 0 & 0.043 \end{pmatrix}$ | $\begin{pmatrix} 0.22 & 0 \\ 0 & 0.059 \end{pmatrix}$ |
| $\hat{\pi}_{11}$ | $\begin{pmatrix} 0.20 & 0 \\ 0 & 0.05 \end{pmatrix}$ | $\begin{pmatrix} 0.19 & 0 \\ 0 & 0.07 \end{pmatrix}$ |
| $\hat{P}_{ij}$ | $\begin{pmatrix} 0.45 & 0.0004 \\ 0.0004 & 0.55 \end{pmatrix}$ | $\begin{pmatrix} 0.55 & 0.0004 \\ 0.0003 & 0.44 \end{pmatrix}$ |

Table 2
Estimated data driven parameters for the non-stationary procedures

| Parameters | P-NSFMC | NP-NSFMC |
|---|---|---|
| $(\hat{\mu}_0, \hat{\mu}_1)$ | (119.99, 142.12) | (120.10, 141.97) |
| $(\hat{\sigma}_0, \hat{\sigma}_1)$ | (4.07, 3.62) | (4.1, 3.72) |

Table 3
Estimated data driven parameters for the stationary procedures

| Parameters | P-SFMC | NP-SFMC | SFMF |
|---|---|---|---|
| $(\hat{\mu}_0, \hat{\mu}_1)$ | (121.02, 141.25) | (120.30, 141.89) | (1119.21, 142.09) |
| $(\hat{\sigma}_0, \hat{\sigma}_1)$ | (4.86, 4.10) | (4.24, 3.76) | (3.75, 3.83) |

Table 4
Rates of error for the stationary and non-stationary procedures

| P-SFMC | NP-SFMC | P-NSFMC | NP-NSFMC | SFMF |
|---|---|---|---|---|
| 5.85% | 6.01% | 5.18% | 5.35% | 6.79% |

The SFMC are staying competitive but do not provide any accuracy information concerning the homogeneity of the image. Actually, the non-stationary procedure estimates correctly the homogeneous areas (see Fig. 3a and b). Let us notice the method using a non-parameterized neighborhood density stay competitive facing the parameterized assumption. Further studies must be performed in order to measure the influence of the parameter $r$.

## 8. Results on real images

We wish to identify different homogeneous regions inside an image. We processed here our images in the mono-spectral context. We present in Fig. 4a and b two images of Oakland typically exhibiting a such situation.

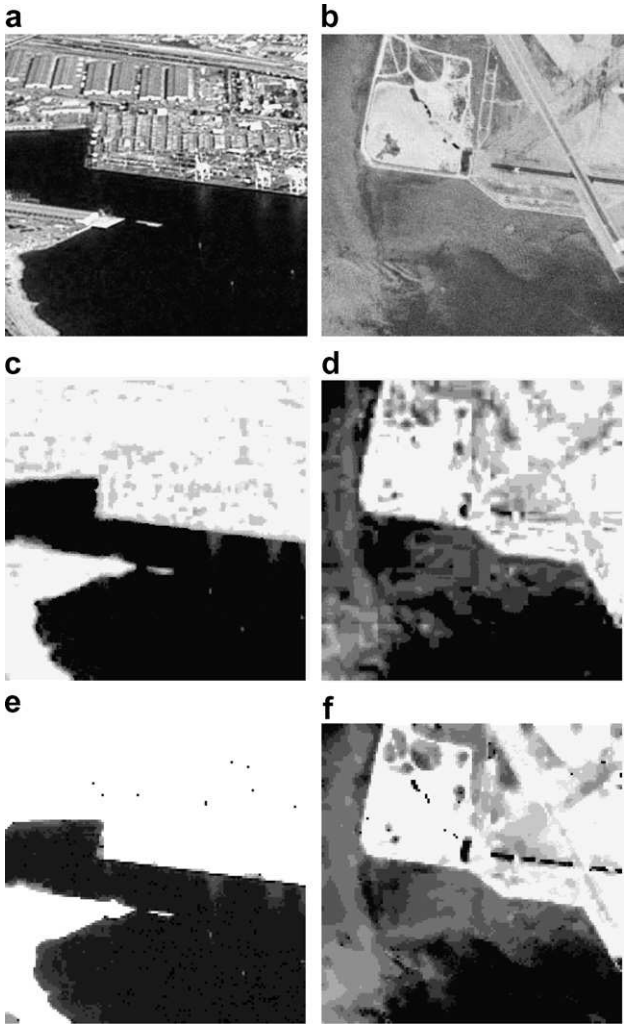



Fig. 4. (a, b) The observation; (c, d) segmented images using SFMC; (e, f) segmented images using SFMF.

Fig. 4a contains a sea area and the city, which appear to be inhomogeneous on the picture, noticing that the distribution of this part of the image behaves differently from the distribution of the sea part. Fig. 4b contains a cloudy area and a town area. We processed first all images Fig. 4a and b using a stationary approach. The results are given in Fig. 4c–f, respectively by the FMC and the FMF algorithm. These methods do not provide the fine details and give comparable results. It is then necessary to take into account the stationary information, initializing the algorithm with more than two stationary states ($K > 2$). The segmented fields $X$ and $U$ are represented, respectively, in Fig. 5a–d. In particular, the class "0" of $U$ indicates the high density region i.e., the city part. Our procedure ensures a convergence of the stationary field towards $K = 2$ classes, which suits the initial hypothesis of both groundtruth models i.e., city/sea areas for Fig. 4a and city/cloud areas for Fig. 4b. The results given by the hidden realizations $X$ are less convenient because the method tends to lose some details, concerning the lower homogeneous regions. In order to enrich the information providing by this field, we propose to combine the NP-FMC and NP-NSFMC approaches into an algorithm as follows:

- (i) Perform the NP-NSFMC method to the observed data.
- (ii) For each stationary state $U_i$, $i = 1, 2, \ldots, K$, apply a stationary NP-FMC method. For each state $i$, the data corresponding to $U_j$, $j \neq i$ are processed as missing data: one has to suppress them in the Hilbert-Peano path (Salzenstein and Collet, 2006).

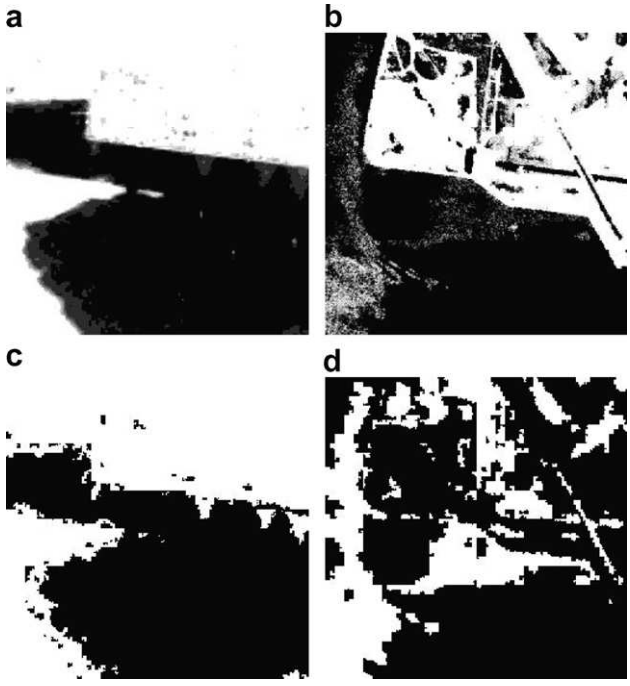



Fig. 5. Segmented images of Fig. 4a and b, using a non-stationary method. (a, b) Realizations of $X$; (c, d) associated $U$ containing two states $K = 2$.

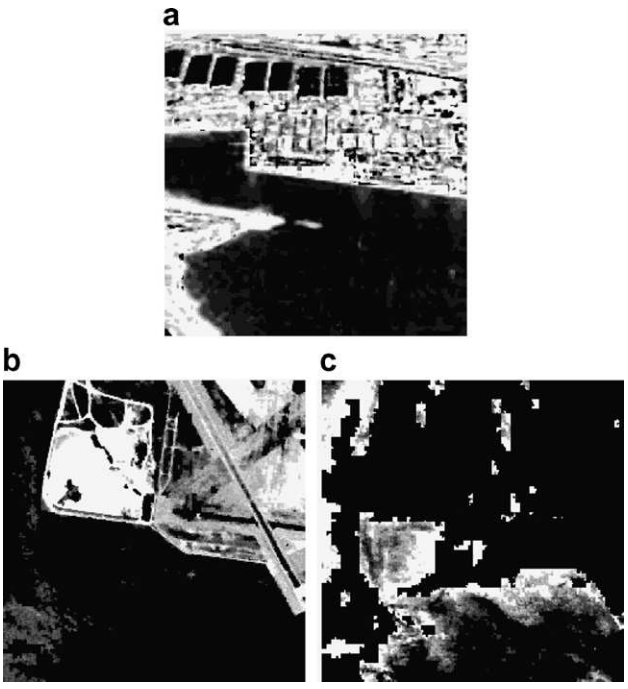



Fig. 6. (a) The segmented areas in Fig. 4a using a two steps non-stationary method. (b) The segmented areas in Fig. 4b using a two steps non-stationary method.

Consequently, we applied the method separately for each stationary parts of Fig. 4a and b. The segmented data corresponding are given in Fig. 6. Although the interpretation is more delicate, requiring separate graylevels/pictures, the final results provide more detailed groundtruth.

## 9. Conclusion

We presented in this paper a new fuzzy Markov chain model based on a non-stationary approach. On one hand we modeled the prior parameters of a stationary chain using a parameterized joint density defined on neighbored sites. On the other hand, we used an intermediate field $U$ in order to govern the switching in the distribution of $X$. Here the classes in $U$ are discrete while the classes in $X$ are continuous. The proposed method merges the fuzzy processing technique and a recent technique that has been used to describe non-stationary images for hard classification. This model is more flexible than the single chain based procedure, in the following manner: (i) the pairwise $(X, U)$ is a Markov chain, but $X$ is not necessary Markovian; (ii) the stationary model is a particular case with one discrete state in $U$. The fuzzy context should be better adapted than the hard approach when the scene owns diffuse structure. In order to take into account the complexity of multiple real situations, it worths to extend this model to the multi-sensors context and non-Gaussian data driven distributions.

## References

Avrachenkov, K., Sanchez, E., 2002. Fuzzy chains and decision-making. Fuzzy Optim. Decis. Making 1 (2), 143–159.

Caillol, H., Hillion, A., Pieczynski, W., 1993. Fuzzy Markov random fields and unsupervised image segmentation. IEEE Trans. Geosci. Remote Sensing 4, 801–810.

Carincotte C., Derrode S., Boucher J.-M., 2004. Chaînes de Markov cachées floues et segmentation non supervisée d'images. In: Logique Floue et Applications (LFA'04). (Nantes), Novembre.

Celeux, G., Diebolt, J., 1985. The SEM algorithm: A probalistic teacher derived from the EM algorithm for the mixture problem. Comput. Statist. 2, 73–82.

Devijver, P., 1985. Baum's forward backward algorithm revisited. Pattern Recognition Lett. 3, 369–373.

Geman, S., Geman, D., 1984. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Trans. Pattern Anal. Machine Intell. 6 (6), 721–741.

Hughes, J., Guttorp, P., Charles, S., 1999. A non-homogeneous hidden Markov model for precipitation occurrence. Appl. Statist. 48, 15–30.

Lanchantin P., Pieczynski W., 2004. Unsupervised non-stationary image segmentation using triplet Markov chains. In: Advanced concepts for intelligent vision systems (ACVIS 04) (Brussel, Belgium), August.

Maroquin, J., Mitte, S., Poggio, T., 1987. Probabilistic solution of ill-posed problems in computational vision. J. Amer. Statist. Assoc. 82, 76–89.

McLachlan, G., Krishnan, T., 1997. EM Algorithm and Extensions. Wiley.

Mohammed, M., Gader, P., 2000. Generalized hidden Markov models – Part I: Theoretical frameworks. IEEE Trans. Fuzzy Syst. 8 (1), 67–81.

Pieczynski, W., 2002. Chaînes de Markov Triplet, Triplet Markov Chains. Acad. Sci. Rep. 335 (3), 275–278.

Salzenstein, F., Collet, C., 2006. Fuzzy Markov random fields versus chains for multispectral image segmentation. IEEE Trans. Pattern Anal. Machine Intell. 11 (28), 1753–1767.

Salzenstein, F., Pieczynski, W., 1997. Parameter estimation in hidden fuzzy Markovian fields and image segmentation. Graphical Models Image Process. 59 (4), 205–220.

COMPUTER METHODS AND PROGRAMS IN BIOMEDICINE 90 (2008) 191–201

journal homepage: www.intl.elsevierhealth.com/journals/cmpb

# Contrast enhancement in emission tomography by way of synergistic PET/CT image combination

N. Boussion[a,*], M. Hatt[a], F. Lamare[b], C. Cheze Le Rest[a], D. Visvikis[a]

[a] INSERM, U650, Laboratoire du traitement de l'information médicale (LaTIM), CHU Morvan, 29609 Brest, France
[b] CSC PET Cardiology, MRC Clinical Sciences Centre, Faculty of Medicine, Imperial College, London, UK

ARTICLE INFO

*Article history:*
Received 16 August 2007
Received in revised form
17 December 2007
Accepted 20 December 2007

*Keywords:*
PET
CT
Image display
Image combination
Wavelets

ABSTRACT

The display of image fusion is well accepted as a powerful tool in visual image analysis and comparison. In clinical practice, this is a mandatory step when studying images from a dual PET/CT scanner. However, the display methods that are implemented on most workstations simply show both images side by side, in separate and synchronized windows. Sometimes images are presented superimposed in a single window, preventing the user from doing quantitative analysis. In this article a new image fusion scheme is presented, allowing performing quantitative analysis directly on the fused images.

*Methods:* The objective is to preserve the functional information provided by PET while incorporating details of higher resolution from the CT image. The process relies on a discrete wavelet-based image merging: both images are decomposed into successive details layers by using the "à trous" transform. This algorithm performs wavelet decomposition of images and provides coarser and coarser spatial resolution versions of them. The high-spatial frequencies of the CT, or details, can be easily obtained at any level of resolution. A simple model is then inferred to compute the lacking details of the PET scan from the high frequency detail layers of the CT. These details are then incorporated in the PET image on a voxel-to-voxel basis, giving the fused PET/CT image.

*Results:* Aside from the expected visual enhancement, quantitative comparison of initial PET and CT images with fused images was performed in 12 patients. The obtained results were in accordance with the objectives of the study, in the sense that the organs' mean intensity in PET was preserved in the fused image.

*Conclusion:* This alternative approach to PET/CT fusion display should be of interest for people interested in a more quantitative aspect of image fusion. The proposed method is actually complementary to more classical visualization tools.



## 1. Introduction

Multimodality imaging has become a mandatory exploration in many clinical applications. PET/CT hybrid scanners constitute today a necessary tool in diagnosis, treatment and staging of cancer [1–3]. The complementary information provided by this kind of dual imaging device allows revealing the physiological state of malignant tumours by PET while in the same time, the CT image offers anatomical accuracy through high-spatial resolution. One of the specific uses of PET/CT that currently encounters increasing interest is intensity-modulated radiotherapy (IMRT). Recent works tend to prove

* *Corresponding author*. Tel.: +33 298018105.
E-mail address: nboussion@yahoo.fr (N. Boussion).
0169-2607/$ – see front matter © 2008 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.cmpb.2007.12.009

indeed that PET/CT-guided IMRT improves treatment planning while reducing tissue doses, for example in head and neck cancer [4].

Aside from the problem of spatial co-registration of PET and CT images, which permits their superposition on a voxel-to-voxel basis, the effective management of images in a day-to-day clinical use consists of their visualization. This step is of great importance since it allows comparing the images and making an accurate judgment according to functional and morphological complementarity. Two main types of visualization techniques exist [5]. In the first one, the two images are displayed side by side in two separate windows, with synchronized commands and cursors. This method of display presents the advantage of preserving information but the efficient visual comparison of structures remains difficult. The second visualization approach is the overlay of both images in a single window. Several approaches can be chosen but most of them require two look-up-tables, generally a grey level one for the CT image and a colour one for the PET image. An easy way to proceed is to display a voxel of each image alternately, like a mosaic. Another approach is to blend the images using a single look-up-table but then the intensity in a voxel is a weighted sum of PET and CT intensities in the given voxel. Recent works in this domain include multi-image voxel composting [6] in which the CT image is decomposed into several layers with different ranges of contrast adjustments, each one corresponding to a particular tissue (bone, lungs, soft tissues, etc.). These layers are then weighted and mixed together, finally been blended with the PET image.

In oncology staging or treatment planning it is of decisive importance to follow the evolution of both tumour activity and size. However, such quantitative measurements cannot be directly derived from the fused images since the intensity in a voxel is a mixture of corresponding PET and CT intensities. Even if the anatomical information is preserved to a certain extent, it is impossible to measure the PET intensity in a given region of interest. Moreover, the complementarity aspect of fusion display loses its interest in the sense that it is limited to visual inspection.

By enlarging the field of investigation, one may notice that image fusion concerns many fields, like geosciences [7], food safety [8], fingerprints analysis [9], biometric imaging [10] or forensic investigations [11]. Nevertheless, the objective of all these studies remains largely in the scope of visual enhancement of images without considering the quantitative aspect which is of outmost importance in medical imaging. Some general surveys have been also performed but mostly tackling without specific attention the medical aspect of image fusion algorithms [12].

In this article, we introduce a new fusion display scheme able to preserve the quantitative functional information provided by PET, and in the same time, able to maintain morphological details of the CT. The algorithm is based on multi-resolution analysis of the PET and CT images using wavelets. After presenting the theory and implementation of this method, we apply it on a number of clinical whole-body example image datasets and perform quantitative analysis to demonstrate its potential for preserving relevant information.

## 2. Background

### 2.1. Basic theory on continuous wavelet transform (CWT)

For the sake of clarity, definitions are given for an 1D function $f$, but a more general theory can be found in [14]. The wavelet transform $W$ of an 1D, real, square-integrable function $f$ is defined by

$$W(a,b) = \int_{-\infty}^{+\infty} f(x)\psi^*\left(\frac{x-b}{a}\right)\,dx,$$

where $a$ is the scale of the analysis and $b$ is the parameter of translation corresponding to the position of the wavelet $\Psi$ ($\Psi^*$ stands for the complex conjugate of $\Psi$). $W(a,b)$ is the inner product of $f$ with the scaled and translated versions of $\Psi$:

$$W(a,b) = \int_{-\infty}^{+\infty} f(x)\psi^*_{a,b}(x)\,dx = < f(x), \Psi_{a,b}(x) >,$$

with

$$\Psi_{a,b}(x) = \left(\frac{1}{\sqrt{a}}\right)\Psi\left(\frac{x-b}{a}\right).$$

$W$ may also be seen as a measure of similarity between the function $f$ and the basis functions $\Psi_{a,b}$ which are derived from the so-called mother wavelet $\Psi$. Here, similarity refers to a comparable frequency content, at the current scale $a$. The wavelet analysis can then be seen as a mathematical microscope which does not depend on the magnification once the optical $\Psi$ is chosen.

The reconstruction formula is

$$f(x) = C_\Psi^{-1}\int_0^{+\infty}\int_{-\infty}^{+\infty}\sqrt{a}W(a,b)\psi(x-b/a)\,da\,db/a^2,$$

with

$$C_\Psi = \int_0^{+\infty}|\hat{\Psi}(\upsilon)|^2\,d\upsilon/\upsilon$$

($\hat{\Psi}$ is the Fourier transform of $\Psi$).

### 2.2. Discrete wavelet transform (DWT)

The discrete wavelets that have been introduced are no longer continuously translatable and scalable but can simply be translated and scaled following discrete steps with indices $j$ and $k$:

$$\Psi_{j,k}(x) = \frac{1}{\sqrt{a^j}}\Psi\left(\frac{x-kba^j}{a^j}\right).$$

This very general solution has been proposed by Daubechies [13], who also suggested the choice of $a=2$ and $b=1$ (dyadic sampling) to obtain orthogonal basis functions for certain $\Psi$ wavelets (Daubechies wavelets). The multi-resolution analysis developed by Mallat [14] allows to implement this discrete approach by using high-pass and low-pass filtering and sub-sampling. However, this algorithm cancels shift invariance which may be a problem when dealing with several images like in image fusion. The resulting wavelet transform is no longer shift invariant, which means

that the wavelet transforms of an image and of a shifted version of the same image are not simply shifted versions of each other. For this reason we have chosen to use the undecimated "à trous" algorithm which is totally shift invariant and also easier to implement. It also has the advantage of leading to wavelet images of same size as the original image, allowing to compare details at different scales on a voxel-to-voxel basis. The general theory linked to this algorithm is extensively presented in [15], where accurate comparison with the Mallat implementation of DWT is also provided.

## 3. Design considerations

In this section we present the iterative "à trous" algorithm which can be easily implemented on a given image $I$. This discrete wavelet transform algorithm was introduced by Dutilleux [16], developed by Holdschneider [17] and detailed by Starck et al. [18]. The process gives an image sequence of coarser and coarser spatial resolution by performing successive convolutions with a low-pass filter $h$. At each iteration $j$, the spatial resolution of the approximation image $\text{app}_{j-1}$ is degraded to give the approximation image $\text{app}_j$ according to

$$\text{app}_j(k,l) = \sum_{m,n} h(m,n)\text{app}_{j-1}(k+m2^{j-1}, l+n2^{j-1}).$$

The first approximation image $\text{app}_0$ is taken as $I$, the original image. The difference $\text{app}_{j-1} - \text{app}_j$ is the wavelet coefficients $w_j$ containing the details (edges, texture) at a resolution level between $\text{app}_{j-1}$ and $\text{app}_j$. The synthesis procedure that reconstructs the original image from its layers of details $w_k$ is given by

$$\text{app}_0 = I = I_N + \sum_{k=1}^{k=N} w_k,$$

with $N$ the number of iterations from the initial image $I$ to the final approximation $I_N$ of spatial resolution decreased by $2^N$.

The algorithm can be easily implemented by performing the following steps [18]:

1. Initialize $j$ to 0: start with the original image $I = \text{app}_0$ (app stands for approximation).
2. Increment $j$ and carry out a convolution of $\text{app}_{j-1}$ with a low-pass filter $h$ in order to obtain $\text{app}_j$ (the distance between the central voxel and the adjacent ones is $2^{j-1}$).
3. The wavelet coefficients $w(j)$ at this level of resolution are given by $\text{app}_{j-1} - \text{app}_j$.
4. If $j$ is less than the required number $N$ of resolutions go to step 2.
5. The set $W = \{w(1), w(2), \ldots, w(N), \text{app}_N\}$ is the wavelet transform of $I$.

Practically, at each iteration, zeros are inserted between lines and columns of the filter $h$ giving its name to the algorithm "à trous" which in French means "with holes". In this work, the chosen low-pass filter $h(k) = (1/16, 1/4, 3/8, 1/4, 1/16)$

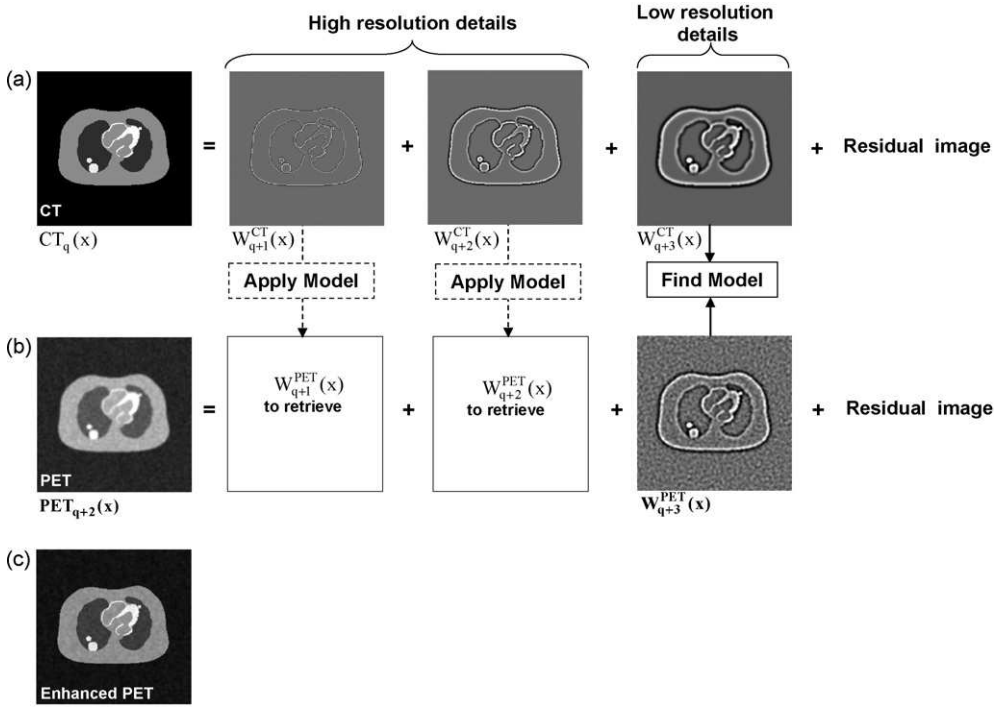



**Fig. 1 – Illustration of the PET enhancement process on simulated images (a) wavelet transform of the original simulated CT image (resolution level $q$); (b) wavelet transform of the original simulated PET image (resolution level $q+2$); the lacking details of the PET image, $W^{PET}_{q+1}(\vec{x})$ and $W^{PET}_{q+2}(\vec{x})$, are retrieved from the existing details $W^{CT}_{q+1}(\vec{x})$ and $W^{CT}_{q+2}(\vec{x})$ of the CT images which are modified according to the model defined between existing details of both CT and PET at a lower but common level of resolution; (c) the enhanced PET image is the voxel-to-voxel addition of the original PET image with the retrieved details $W^{PET}_{q+1}(\vec{x})$ and $W^{PET}_{q+2}(\vec{x})$.**

is deduced from the spline function of order 3 [19]:

$$\phi(x) = \frac{1}{12}(|x-2|^3 - 4|x-1|^3 + 6|x|^3 - 4|x+1|^3 + |x-2|^3).$$

In 2D we have $\phi_{2D}(x, y) = \phi(x)\phi(y)$ thus leading to $h_{2D}(k, l) = h(k)h(l)$, the filter being separable:

$$h_{2D}(k,l) = \begin{bmatrix} 1/256 & 1/64 & 3/128 & 1/64 & 1/256 \\ 1/64 & 1/16 & 3/32 & 1/16 & 1/64 \\ 3/128 & 3/32 & 9/64 & 3/32 & 3/128 \\ 1/64 & 1/16 & 3/32 & 1/16 & 1/64 \\ 1/256 & 1/64 & 3/128 & 1/64 & 1/256 \end{bmatrix}.$$

This filter is isotropic, has a Gaussian-like shape and can be convolved in two steps (columns and then lines).

## 4. System description

### 4.1. *Alternative "à trous" implementation and combination process*

The method relies on the fact that PET and CT images are spatially co-registered, i.e. both images can be superimposed and are reconstructed with the same voxel size. The fusion display approach presented here aims at preserving relevant information provided by each modality: anatomical details and high-spatial resolution from CT on the one hand, and functional data from the PET on the other hand. For this purpose the anatomical details provided by the CT and corresponding to resolution levels that are not present in the PET are detected, extracted, modified and injected in the PET image on a voxel-to-voxel basis.

Wavelet analysis allows the spatial frequencies to be easily obtained, in particular at a level of resolution common to the CT and PET images. A model is then inferred to estimate the lacking details of PET from the high frequency details layers of CT. If the level of resolution of CT is $q$, referred to as $CT_q$, and the one of PET is $r = q + p$, referred to as $PET_r$, we can write

$$PET_r(\vec{r}) = PET_{q+p}(\vec{r}) = PET_{q+p+1}(\vec{r}) + w^{PET}_{q+p+1}(\vec{r}),$$

and

$$CT_q(\vec{r}) = CT_{q+p+1}(\vec{r}) + \sum_{k=1}^{k=p+1} w^{CT}_{q+k}(\vec{r}).$$

The lacking details of PET are the wavelet coefficients $w^{PET}_i(\vec{r})$ with $q \le i \le q+p$; however we do possess $w^{PET}_{q+p+1}(\vec{r})$ and $w^{CT}_{q+p+1}(\vec{r})$ and we assume that there exists a more or less simple link between them like $w^{PET}_{q+p+1}(\vec{r}) = \alpha \times w^{CT}_{q+p+1}(\vec{r}), \alpha \in IR^*$ for instance. Although, different models can be envisaged, in this study a simple linear model is used

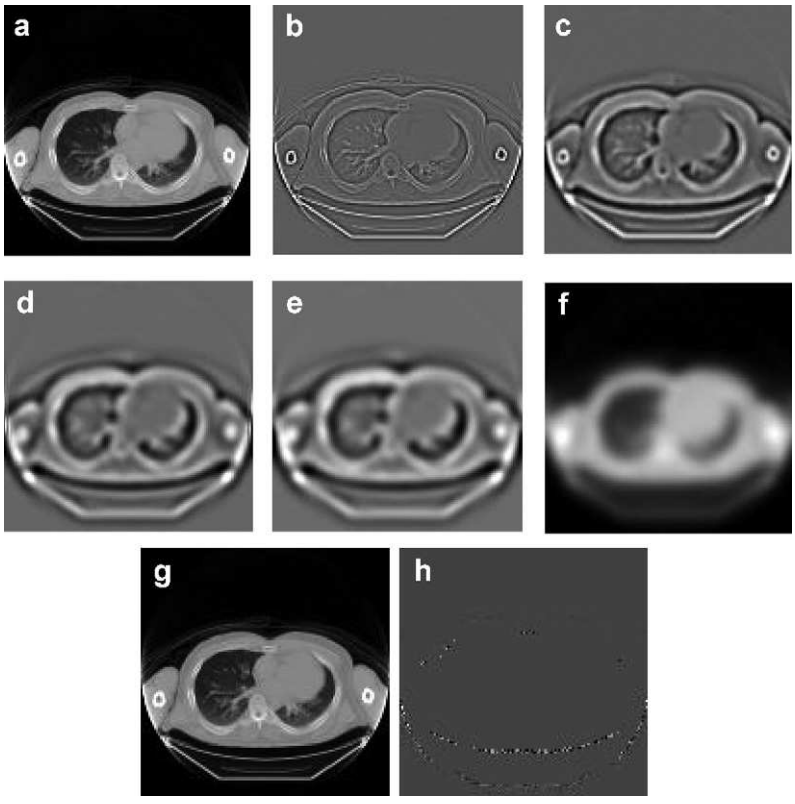



**Fig. 2 – An example of CT decomposition using dyadic and linear transformation ("à trous" algorithm). (a) original CT image of resolution 1 mm; (b)-(e) wavelet layers corresponding to details of resolution 1 mm-2 mm, 2 mm-4 mm, 4 mm-5 mm and 5 mm-6 mm, respectively; (f) residual approximation image; (g) reconstructed CT image corresponding to the pixel-to-pixel addition of wavelet layers (b)–(e) and residual image (f); (h)difference between original CT image (a) and reconstructed CT image (g).**

where the parameter alpha is considered equal to the mean voxel-to-voxel division of $w^{PET}_{q+p+1}(\vec{r})$ by $w^{CT}_{q+p+1}(\vec{r})$.

$PET_q$ can now be reconstructed from $PET_r$ by taking the $w^{PET}_i$ $(q \le i \le q+p)$ into account. They are calculated as $w^{PET}_i(\vec{r}) = \alpha \times w^{CT}_i(\vec{r})$, $(q \le i \le q+p)$ leading to

$$PET_q(\vec{r}) = PET_{q+p+1}(\vec{r}) + \sum_{k=1}^{k=p+1} w^{PET}_{q+k}(\vec{r}).$$

As stated above, the undecimated "à trous" implementation of the DWT is dyadic which means that the resolution decreases by a factor of two at each iteration. Thus if the resolution of CT is 1 mm and the one of PET is 8 mm the process is able to separate the wavelet images of CT in the sequence $w_{1-2mm}$, $w_{2-4mm}$, and $w_{4-8mm}$, and the resulting CT approximation has a resolution of 8 mm, exactly the same as the PET. However, if the resolution of PET is 6 mm, it is impossible to obtain a CT approximation with a resolution equal to 6 mm. The resolutions that can be obtained are only 2 mm, 4 mm, 8 mm, 16 mm and so on. For this reason we modified the classical implementation of the algorithm to make any discrete resolution level available. In the normal implementation, the sampling of the image to be convolved is widened by a factor of 2 at each iteration. This is done indirectly by inserting zeros in the filter mask: the distance between the studied voxel and its neighbours is $2^{j-1}$. This sampling scheme is manda-

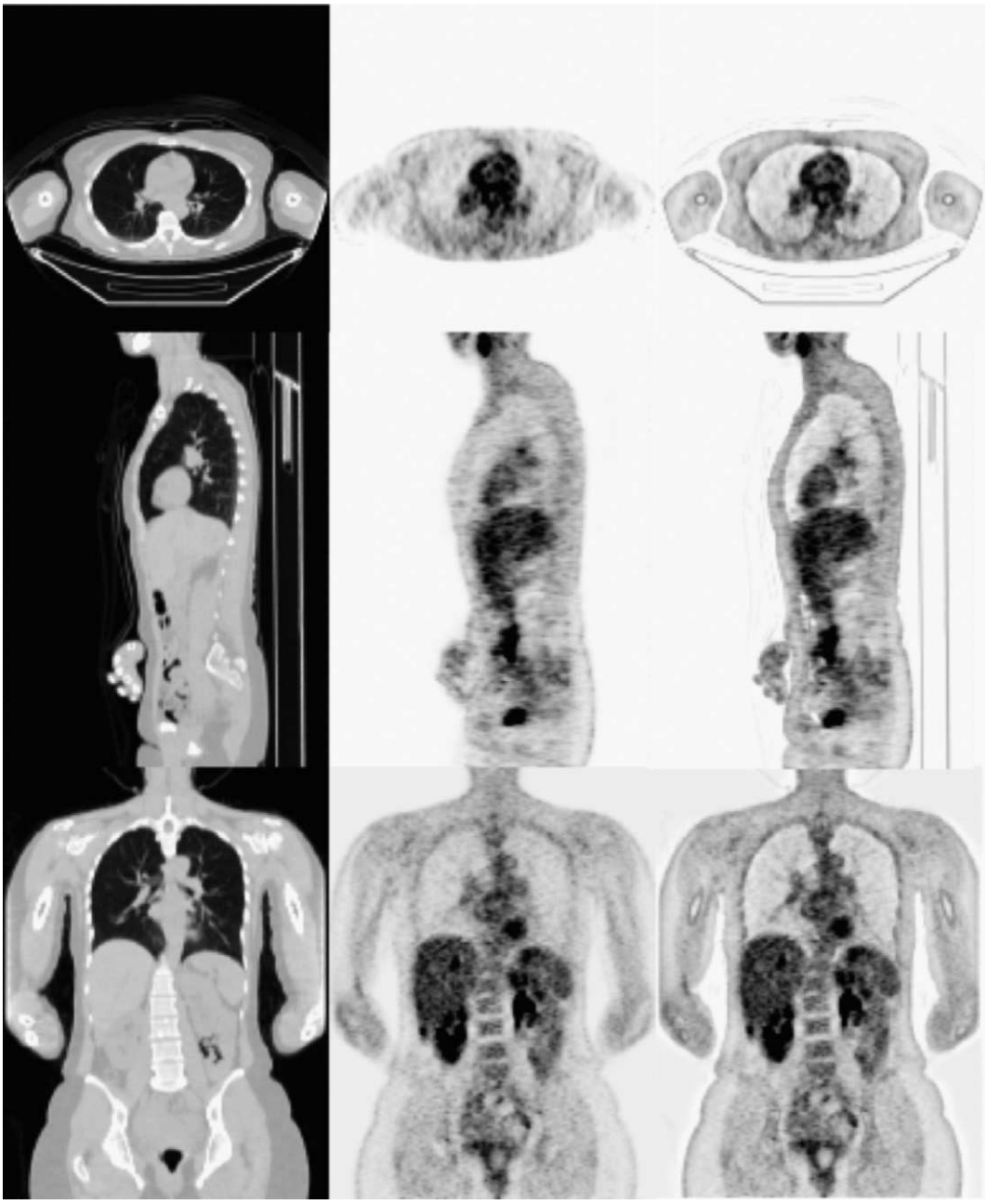


**Fig. 3 – Example of PET contrast enhancement by combination with CT. First column: CT images (transverse, sagittal, coronal); second column: PET images; third column: enhanced PET images. The combination scheme was dyadic + linear decomposition.**

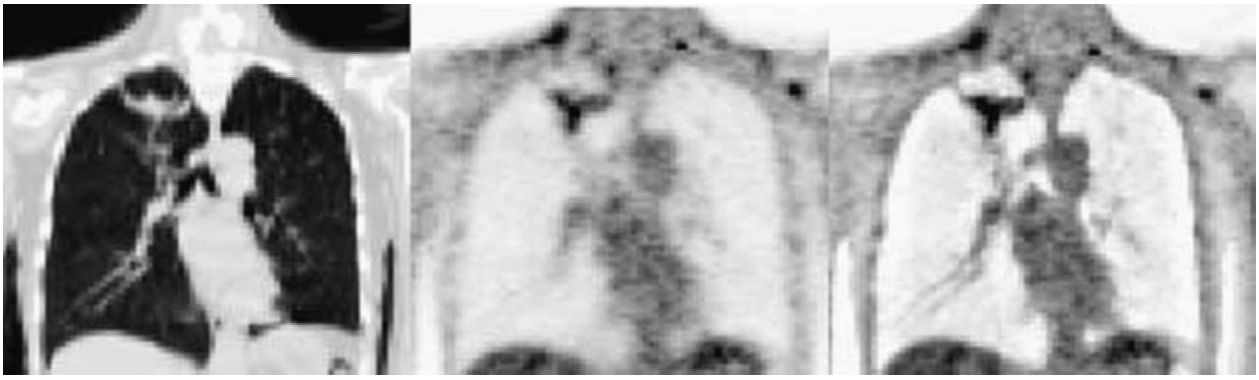


**Fig. 4 – Supplementary example of fusion, here on a restricted area corresponding to lungs (Patient 6). From left to right: CT, PET, enhanced PET.**

tory to get perfect reconstruction, but the CT reconstruction is not our final objective. The aim is to extract wavelet layers of details corresponding to resolution levels higher than the resolution of the PET image. Consequently, we modified the "à trous" algorithm to get a linear version instead of a dyadic one. The sampling is then performed differently in the sense that zeros are still inserted in the filter, but the distance between the centre voxel and its neighbours becomes $j$ instead of $2^{j-1}$. The obtained series of resolution levels is then 2 mm, 3 mm, 4 mm, 5 mm and so on. From this alternative implementation one can also deduce a third approach as a mixture of the two and referred to as "dyadic + linear". In this method, the dyadic implementation is performed first, leading to the series of resolutions 2 mm, 4 mm for instance, and then the linear approach is performed giving resolutions of 5 mm, 6 mm, etc. The main interest consists of a gain in computation time in comparison with a linear implementation alone.

In this study, the model between details of PET and CT at their common level of resolution is the mean voxel-to-voxel division of the layer images containing wavelet coefficients. The obtained parameter is then multiplied to each wavelet

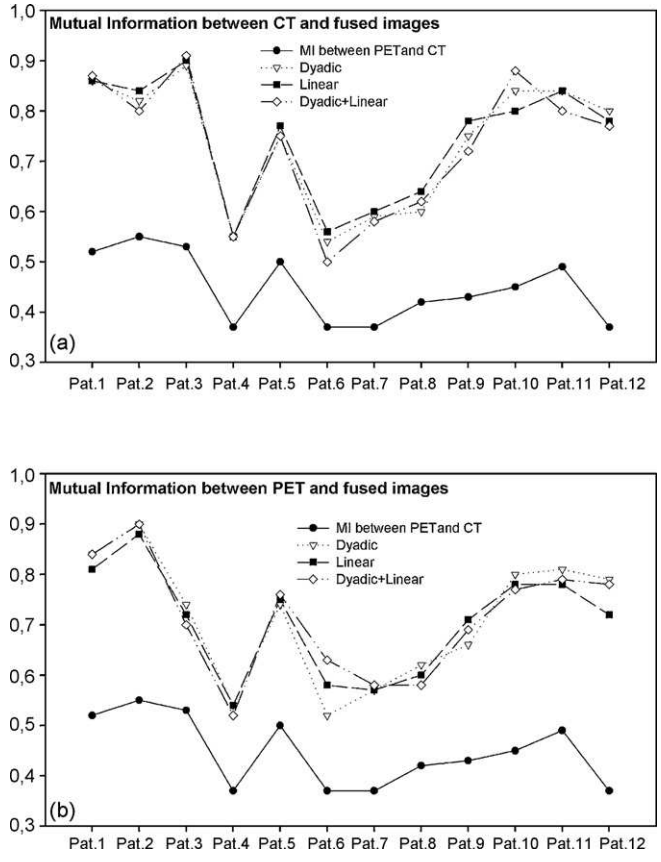



**Fig. 5 – Mutual information (MI) between initial and fused images, for the three presented methods (dyadic, linear, dyadic + linear). Left: MI between CT and PET/CT fused images. Right: MI between PET and PET/CT fused images.**

image (always on a voxel-to-voxel basis) of the first CT decomposition. These modified images containing high-resolution details not present in the PET are added to the said PET giving the actual fused image. A graphical description of the whole process is shown in Fig. 1 in order to facilitate the understanding of the method.

### 4.2. *Clinical images and quantitative assessment*

The presented algorithm and its different implementations were tested on 12 clinical images. They consisted of whole-body FDG PET and CT images of patients under oncological follow-up. Images were obtained on a dedicated PET/CT scanner (GE DLS) using a 5-min per bed position for the emission acquisition (2D mode) and a CT scan (140 kV, 80 mA) over the same area acquired under shallow breathing. Images were reconstructed using ordered subsets expectation maximisation (OSEM) with two iterations and 28 subsets. The CT maps were used for attenuation correction [19] after being reduced to $128 \times 128$ with a reconstruction slice interval of 4.25 to match the in-slice resolution and the slice thickness of the PET reconstructed images.

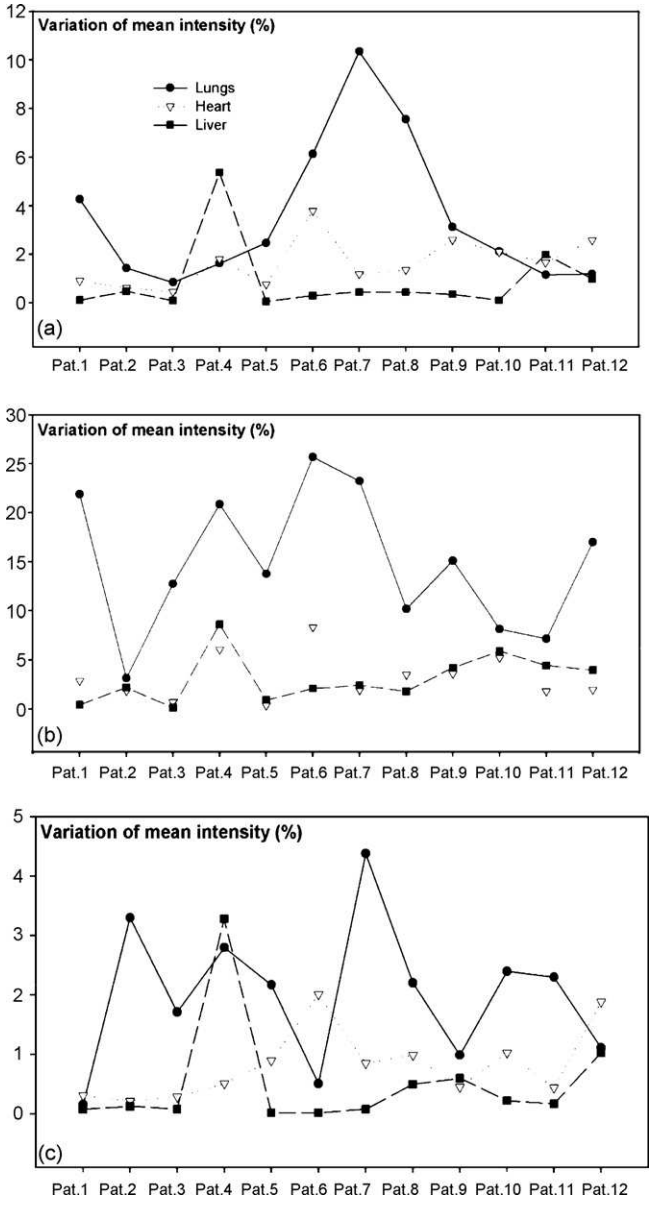



**Fig. 6 – Percentage of intensity variation between PET and fused PET/CT in a series of ROIs (lungs, heart, liver). (a) Dyadic analysis; (b) linear analysis; (c) dyadic + linear.**

To our knowledge, there exists no "gold standard" approach for evaluating a fusion display apart from visual assessment. In this study considering the proposed methodology we have chosen quantitative indices to evaluate the algorithm performance and more particularly to assess its specificity regarding preservation of PET information. We have therefore compared the intensity in regions of interest (ROI) corresponding to different tissues, in the PET images and then in the fused one, for each patient. The regions investigated were lungs, liver and heart and the ROIs were drawn manually on the PET and reported automatically onto the fused image. Mean signal and standard deviation in these ROIs were then calculated and compared.

On the other hand in order to evaluate the fusion aspect of the display we calculated also the mutual information between PET and CT on the one hand and PET and the fused image on the other. Given two images $M$ and $N$, their mutual information is given by

$$I(M,N) = \sum_{m \in M} \sum_{n \in N} p(m,n) \log \frac{p(m,n)}{p(m)p(n)}$$

where, $p(M, N)$ is the joint histogram of $M$ and $N$ and $p(M)$ and $p(N)$ the histograms of $M$ and $N$, respectively. $I(M, N)$ can be seen as the amount of information that is common to $M$ and $N$. Consequently we calculated $I$(PET,CT), $I$(PET, fused) and $I$(CT, fused).

As a last part, we calculated the mean contrast along linear plot profiles placed at two kinds of tissue boundaries. The first group of profiles was put at the interface between lungs and soft tissue (liver or heart) and the second group consisted of profiles put across small tumours (10–40 mm) inside homogeneous regions, mainly lungs and liver. In each obtained profile the contrast was calculated using the following formula:

$$\text{contrast} = 100 \times \frac{|xj - xi|}{xj + xi},$$

where $xi$ and $xj$ are the values of two adjacent pixels along the slope of the profile. The local contrast was calculated on each pixel of the slope and the mean contrast was the mean of these values.

## 5. Status report

An example of CT image decomposition using the "à trous" algorithm (version using dyadic then linear transformation) is shown in Fig. 2. The reconstructed CT image (Fig. 2g) is very similar to the original one (Fig. 2a) since their voxel-to-voxel difference (Fig. 2h) has only zero values apart from a limited number of voxels. The quantitative measurements on Fig. 2h give mean value $1.9 \times 10^{-10} \pm 7.4 \times 10^{-7}$, min value $-1.5 \times 10^{-5}$ and max value $1.5 \times 10^{-5}$ (mean of the absolute values $5.4 \times 10^{-6} \pm 5.7 \times 10^{-6}$) while the same investigation when using only dyadic decomposition would lead to a zero

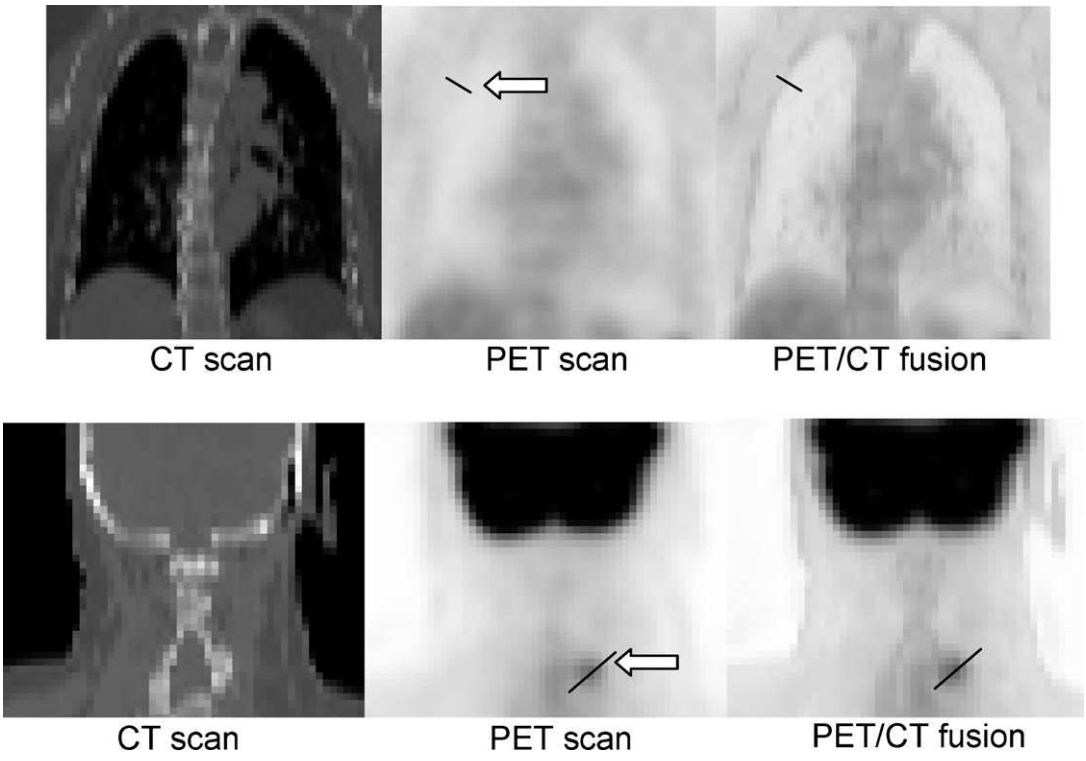



**Fig. 7 – Example of PET/CT fusion (Patient 10) and location of two lines of interest (arrows) for calculating contrast. Intensity profiles along these two lines are given in Fig. 8. Top: lungs-soft tissue; bottom: isolated tumour.**

mean exactly. Fig. 3 shows the fusion results for Patient 5, where CT and FDG PET images are illustrated as well as their fusion, in the three planes of space (transverse, coronal, sagittal). Patient 6 is shown in Fig. 4 where one can see with more details the upper thoracic part of the body corresponding to the lungs. Visually, the fusion allows a contrast improvement of the PET image with details and boundaries better delineated. In the same time, while detailed structures of the CT appear in the fusion image the global intensity of PET is preserved.

Quantitative results are given in Figs. 5 and 6. In Fig. 5 one can see the values of mutual information between PET, CT and fused PET/CT providing a degree of similarity between images. The curves' shapes in Fig. 5a and b are similar and show that the degree of similarity between PET or CT and PET/CT fused images is much larger than the one between PET and CT (MI mean increase 87%). A number of points can be made based on this result; namely that both initial images have contributed to the fused image, and that the fused image contains more information than the PET alone and the CT alone, which is in accordance with the aim of a fusion process. On the other hand, still considering the mutual information, it appears that the three methods of analysis presented in Section 4.1, that is, dyadic, linear, and dyadic + linear gave comparable similarity values between images.

The variation of intensity between PET and fused PET/CT images are given in Fig. 6 in which results are given also for the three proposed methods (dyadic, linear and dyadic + linear). One can see that concerning heart and liver, the difference never reaches 6% for the dyadic and dyadic + linear approaches, while it remains under 10% for the linear approach. As far as the lung results are concerned, the difference in intensity between PET and fused PET/CT is larger, attaining 25% for Patient 6 when using the linear approach (Fig. 6b). However, the mean activity in lungs stays relatively stable (variation less than 10%) when using the dyadic method (Fig. 6a), and stays below 5% of variation when using the dyadic + linear analysis (Fig. 6c).

Plot profiles along lines crossing tissue boundaries were used to qualitatively evaluate the gain in contrast. Two examples are given in Figs. 7 and 8 where lines of interest are given (for Patient 10) as well as their corresponding intensity profiles. One can notice that the slope of the curves is increased after PET/CT fusion (Fig. 8), proving that not only the PET image is improved visually, but also in the same time contrast at boundaries is increased. Quantitatively, the mean contrast at boundaries between large regions (lungs–soft tissues) was increased by +74.5% ±21.4% (min 36.1%, max 88.8%, three profiles by patient, total 36 profiles). In the set of profiles across isolated tumours inside homogeneous areas, the contrast increase was +52.5% ±19.5% (min 29.1%, max 81.3%, 27 tumours in total).

## 6. Lessons learned

In this article a new approach to PET/CT image fusion has been proposed for whole-body imaging. Contrary to the great majority of existing methods, the aim of the presented work was to provide the user with a fused image preserving both

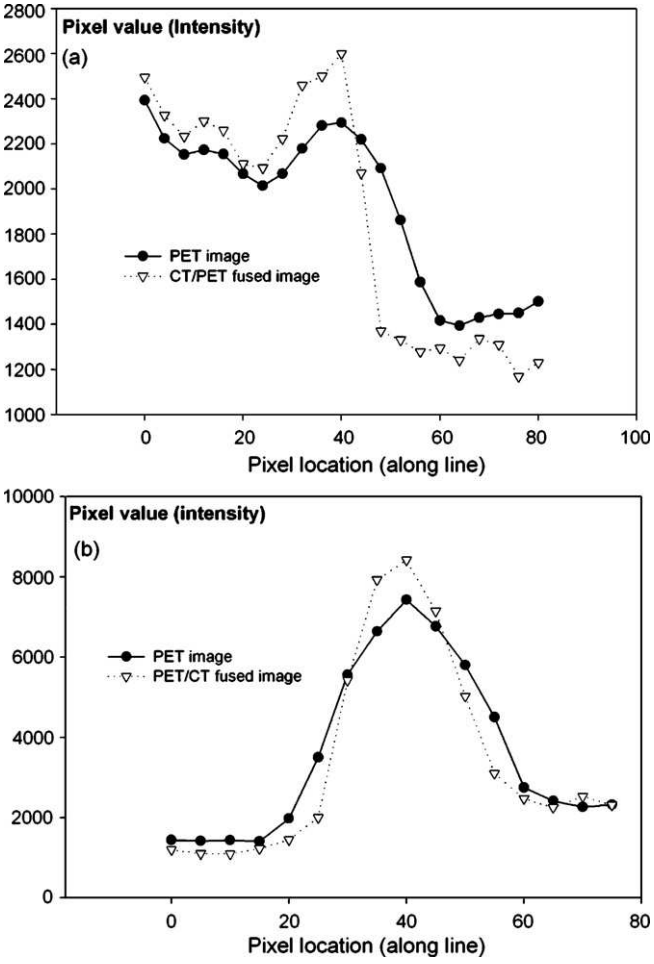



**Fig. 8 – Intensity profiles along segments of line shown in Fig. 7 (Patient 10). After PET/CT fusion, one can observe that slopes are sharper, which corresponds to an improvement of contrast. At lungs–soft tissue boundaries; (b) across tumour of the neck.**

anatomical and functional data. The objective is therefore different from simply presenting two images in a visually convenient fusion display in the sense that quantitative analysis is also here considered as a possible step. In the proposed methodology the anatomical information is present in terms of improved contrast while the intensity in the organs is comparable with the functional information presented by the PET. This is of paramount importance in cancer staging and treatment follow-up for instance, where quantitative assessment of activity uptake is necessary. Furthermore, when using PET/CT-guided IMRT the complementarity provided by the image fusion proposed in this article may be of key interest. Indeed, when considering day-to-day clinical use, this algorithm is more user-friendly and allows physicians to gain a lot of time when making diagnosis and treatment planning for radiotherapy.

The method here described is rather simple to implement and consists of introducing the high-resolution details of the CT in the PET image by using wavelet transform. The discrete implementation of the wavelet transform is performed using original and alternative versions of the "à trous" algorithm. Actually, since the reconstruction of images is less important in this application relative to accessing the wavelet images corresponding to details, we modified the original algorithm in such a way that all levels of resolution are accessible. This alteration allows degrading the CT image to the exact level of the PET resolution. Unfortunately, the discrete wavelet transform prevents us from using real resolution values that are more precise than integer values. Even if satisfying results were obtained in this study, it would be valuable in future work to take into account actual resolution values.

Concerning the different versions of the "à trous" algorithm that are presented, it appears that the "dyadic + linear" approach gives the best quantitative results. As a matter of fact, the mean signal in the lungs before and after fusion seems to change less than with the other two approaches (dyadic alone or linear alone). Nevertheless, it is worth mentioning that in the present study the ratio between spatial resolutions (CT and PET) was 6:1. Consequently, for the dyadic approach, it was impossible to obtain CT details at the resolution 6 mm and we had to stop the decomposition at a level of resolution equal to 8 mm. This point may of course explain why dyadic results were not as good as those obtained with the dyadic + linear approach.

The change in intensity in the lungs could be explained by the incorporation of the pulmonary vasculature present in the CT and not in the PET (Fig. 9). Also, the model that is used to modify the detail layers (images of wavelet coefficient) obtained in the CT decomposition may be considered too simplistic (mean voxel-to-voxel division). A more sophisticated model would be preferable and one can consider that adopting a local model may be more appropriate, in particular for limiting artifacts coming from structures present in the CT only.

The methodology itself is of course not restricted to whole-body imaging in the oncology domain. Provided two co-registered images are available, one functional and the other anatomical, the process can be applied in a wide range of clinical areas. As an illustration (Fig. 10) we provide an example of contrast enhancement of an FDG brain image provided by the fusion with the corresponding T1-weighted MRI. Similar observations made on the whole body images results can be also made in this particular case. Contrast is significantly improved in the enhanced FDG PET, particularly at the interface between white matter and gray matter. Furthermore, mean activity in gray matter is preserved in such a way that further image processing could be performed.

For interested readers, the algorithm is available upon simple request by e-mailing the corresponding author. The program is written in JAVA and can easily be added to the free ImageJ software as an independent plug-in with simple graphical user interface.

## 7. Future plans

Concerning the algorithm, it would be of great interest to define a local model instead of a global one in order to modify the detail layers of the CT. This improvement could lead to the elimination of artifacts corresponding to structures present in the CT but not in the PET. The results presented in this

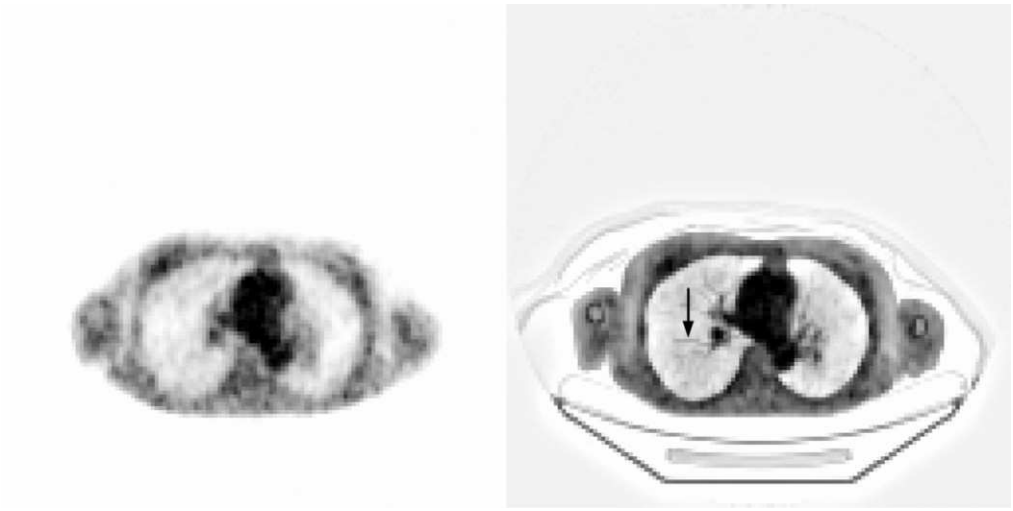


**Fig. 9 – Potential influence of the pulmonary "network". Left: PET before fusion; right: PET enhanced by integration of CT details. The arrow shows that small structures like bronchioles present in the CT may locally but significantly alter the intensity in lungs after combination with PET.**

study were obtained using 2D calculation only. Indeed, most discrete wavelet transforms still perform in 2D and 3D implementations do not exist or are not well validated yet. However, a potential improvement could consist in applying the 2D approach to each plane (sagittal, coronal and transverse) and then making an average from the three obtained images.

To conclude, we have defined a new approach to medical image fusion that allows performing quantitative analysis by means of multi-resolution analysis. The intensity of voxels in the PET image is indeed globally preserved while details of high resolution coming from CT are introduced. As a result, boundaries are better delineated and contrast is enhanced. A

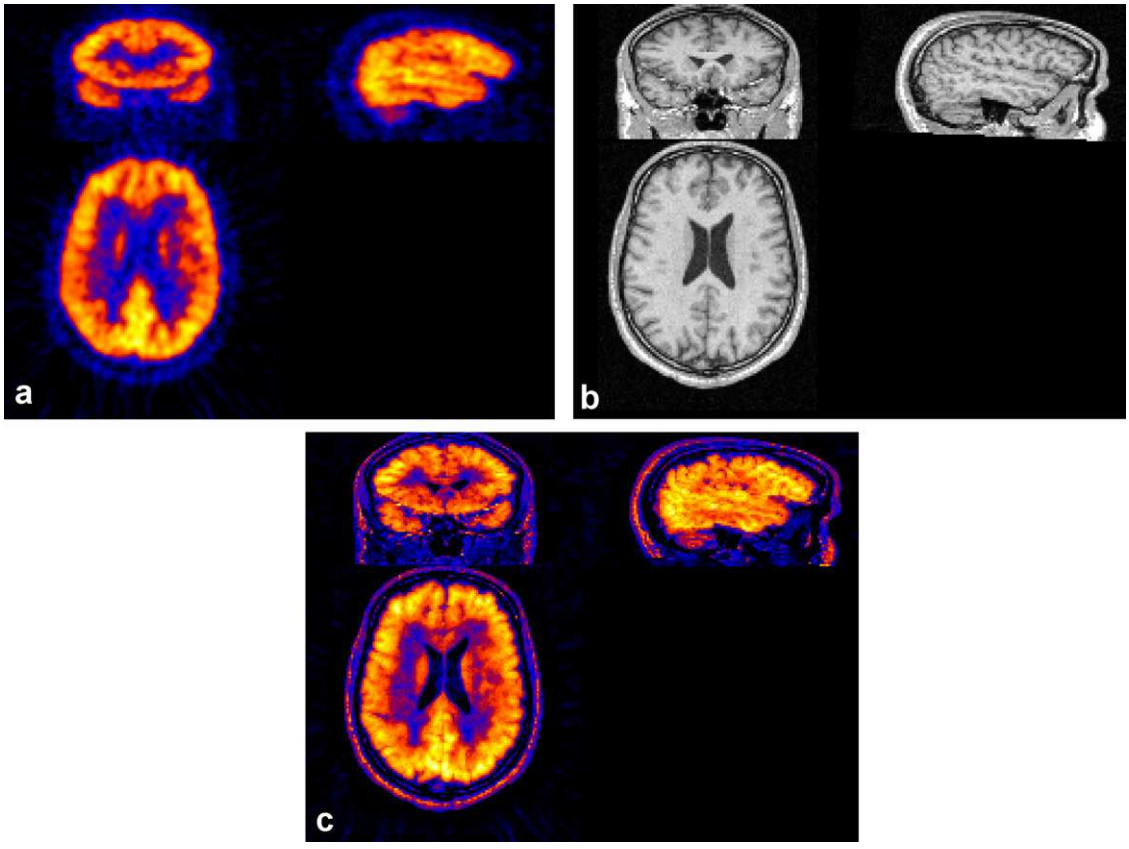



**Fig. 10 – Application of the proposed image enhancement in the cerebral context. Brain FDG PET image; (b) corresponding T1-weighted MRI; (c) enhanced FDG PET after fusion of MRI.**

larger and more comprehensive clinical evaluation of the proposed method should be considered in order to evaluate the potential impact of such a method, for instance in the staging of cancer or treatment planning.

## 8. Conflict of interest statement

None declared.

## REFERENCES


[1] H. Schöder, S.M. Larson, H.W.D. Yeung, PET/CT in oncology: integration into clinical management of lymphoma, melanoma and gastrointestinal malignancies, J. Nuclear Med. 45 (2004) 72S–81S.

[2] H.N. Wagner, Horizons near and far: molecular focus, global involvement, J. Nuclear Med. 46 (2005) 11N–44N.

[3] S. Senan, D. De Ruysscher, Critical review of PET-CT for radiotherapy planning in lung cancer, Crit. Rev. Oncol. Hematol. 56 (2005) 345–351.

[4] D.L. Schwartz, E.C. Ford, J. Rajendran, B. Yueh, M.D. Coltrera, J. Virgin, Y. Anzai, D. Haynor, B. Lewellen, D. Mattes, P. Kinahan, J. Meyer, M. Phillips, M. Leblanc, K. Krohn, J. Eary, G.E. Laramore, FDG-PET/CT-guided intensity modulated head and neck radiotherapy: a pilot investigation, Head Neck 27 (2005) 478–487.

[5] P.J. Slomka, Software approach to merging molecular with anatomic information, J. Nuclear Med. 45 (2004) 36S–45S.

[6] C. Chang, J. Kim, D. Feng, W. Cai, Interactive fusion and contrast enhancement for whole body PET/CT data using multi-image pixel composting, IEEE Nuclear Sci. Symp. Med. Imaging Congress Rec. (2005) 2618–2621.

[7] N.D.S. Gani, M.G. Abdelsalam, Remote sensing analysis of the Gorge of the Nile, Ethiopia with emphasis on Dejen-Gohatsion region, J. African Earth Sci. 44 (2006) 135–150.

[8] M.S. Kim, A.M. Lefcourt, Y.U. Chen, Y. Tao, Automated detection of fecal contamination of apples based on multispectral fluorescence images fusion, J. Food Eng. 71 (2005) 85–91.

[9] G.L. Marcialis, F. Roli, Fingerprints verification by fusion of optical and capacitive sensors, Pattern Recogn. Lett. 25 (2004) 1315–1322.

[10] A. Noore, R. Singh, M. Vatsa, Robust memory-efficient data level information fusion of multi-modal biometric images, Inform. Fusion 8 (2007) 337–346.

[11] F. Puente Leon, Automated comparison of firearm bullets, Forensic Sci. Int. 156 (2006) 40–50.

[12] G. Piella, A general framework for multiresolution image fusion: from pixels to regions, Inform. Fusion 4 (2003) 256–280.

[13] I. Daubechies, Ten Lectures on Wavelets, Presented at CBMS-NSF Regional Conference Series in Applied Mathematics (SIAM), Philadelphia, 1992.

[14] S. Mallat, A theory for multiresolution signal decomposition: the wavelet representation, IEEE Trans. Pattern Anal. Math. Intel. 11 (1989) 674–693.

[15] M.J. Shensa, the discrete wavelet transform: wedding the à trous and Mallat algorithms, IEEE Trans. Signal Process. 40 (1992) 2464–2482.

[16] P. Dutilleux, An Implementation of the Algorithme “à trous” to Compute the Wavelet Transform, Presented at Congrès ondelettes et méthodes temps-fréquence et espace des phases, Marseille, France, Springer–Verlag, 1987, pp. 298–304.

[17] R. Holdschneider, R. Kronland-Martinet, J. Morlet, P. Tchamitchian, A real time algorithm for signal analysis with the help of the wavelet transform, in: J.M. Combes, et al. (Eds.), Wavelets, Springer–Verlag, Berlin, 1989.

[18] J.L. Starck, F. Murtagh, A. Bijaoui, Image processing and data analysis: the multiscale approach, Cambridge University Press, Cambridge, UK, 1998.

[19] J.L. Starck, F. Murtagh, Astronomical Image and Data Analysis, Springer–Verlag, Berlin, 2002.

Eur J Nucl Med Mol Imaging
DOI 10.1007/s00259-009-1065-5

ORIGINAL ARTICLE

# Incorporation of wavelet-based denoising in iterative deconvolution for partial volume correction in whole-body PET imaging

**N. Boussion · C. Cheze Le Rest · M. Hatt · D. Visvikis**

Received: 7 October 2008 / Accepted: 30 December 2008


**Abstract**
*Purpose* Partial volume effects (PVEs) are consequences of the limited resolution of emission tomography. The aim of the present study was to compare two new voxel-wise PVE correction algorithms based on deconvolution and wavelet-based denoising.
*Materials and methods* Deconvolution was performed using the Lucy-Richardson and the Van-Cittert algorithms. Both of these methods were tested using simulated and real FDG PET images. Wavelet-based denoising was incorporated into the process in order to eliminate the noise observed in classical deconvolution methods.
*Results* Both deconvolution approaches led to significant intensity recovery, but the Van-Cittert algorithm provided images of inferior qualitative appearance. Furthermore, this method added massive levels of noise, even with the associated use of wavelet-denoising. On the other hand, the Lucy-Richardson algorithm combined with the same denoising process gave the best compromise between intensity recovery, noise attenuation and qualitative aspect of the images.
*Conclusion* The appropriate combination of deconvolution and wavelet-based denoising is an efficient method for reducing PVEs in emission tomography.

**Keywords** FDG-PET · Image processing · Partial volume correction · Whole-body PET

N. Boussion (✉) · C. Cheze Le Rest · M. Hatt · D. Visvikis
INSERM, U650, Laboratoire de Traitement de l'Information Médicale (LaTIM) CHU MORVAN,
Bat 2bis (I3S), 5 avenue Foch,
Brest 29609, France
e-mail: nboussion@yahoo.fr

## Introduction

Partial volume effects (PVEs) in emission tomography are a well-known consequence of the limited spatial resolution affecting both qualitative and quantitative accuracy of the image. In this work, we refer to spatial resolution as the ability to separate two small objects and, more precisely, the full-width at half-maximum of the point spread function (PSF). PVEs can be divided into three classes; namely (1) attenuation of intensity in small structures compared with the PSF size, (2) spill-in and spill-out affecting bordering organs or, more simply, reciprocal intensity spread from one tissue to another, and (3) tissue mixing in boundaries due to the discrete sampling of images into finite voxels. The last one of these three points is very general as it is a problem in a wide range of medical imaging modalities, both functional or morphological. The voxel size is indeed a physical limit no modality can overcome, whatever its intrinsic spatial resolution. On the other hand, the first two points are more specifically related to quantitative and functional imaging in general and to emission tomography in particular. For this reason, in the present work, we focus only on these two effects.

There exists a wide variety of algorithms aiming at correcting PVEs in emission tomography the vast majority of which have been evaluated in both single photon emission tomography (SPECT) and positron emission tomography (PET). Most of them are based on the use of a priori information provided either by MRI or CT, and remain, even in recent works, restricted to cerebral applications [1–4]. One of these methods, described by Rousset et al. [5], allows an estimation of the real mean signal in any homogeneous tissue providing that its true boundaries are known. This can be achieved by a manual contouring or automatic segmentation (mostly used in brain imaging) of the tissue of interest on the CT or MR image

Published online: 18 February 2009

[6]. However, the assumption related to the tissue homogeneity in the region of interest (ROI) remains questionable and relies mostly on visual analysis of the emission tomography images. Furthermore, adequate tissue correlation between SPECT/PET and MRI or CT cannot be achieved in certain circumstances because of physiological motion in the patient as well as uncontrollable movement of the patient in a more general sense between and during image acquisitions. This is mostly the case in oncology, where images are acquired across the whole body. Unfortunately in this domain structures of interest can be small (tumours) and therefore PVEs significantly alter quantification. For this reason the willingness to incorporate effective PVE correction into PET/CT scanners dedicated to oncology remains of high interest as recently discussed by Basu and Alavi [7]. Furthermore, problems in the use of anatomical information for PVE correction include differences in the appearance and size of some tumours (or structures in general) between the functional and the anatomical images (for example functional necrosis seen on PET may not necessarily be observed on CT). This challenging problem has been widely discussed recently [8].

In an attempt to obviate the need for drawing or segmenting ROIs as well as producing PVE-corrected images that can be used for further processing, we have previously proposed and described a novel voxel-wise PVE correction method for a broad range of clinical applications [9]. On the other hand, this methodology involves mutual analysis of PET and a coregistered anatomical image, and consequently suffers from the limitations related to the tissue spatial correlation already described. In a couple of recent papers, alternative approaches to PVE correction in PET using iterative deconvolution of emission data only have been proposed. The first study was restricted to the cerebral domain [10] and in the second the method was ROI-based only [11]. An interesting third approach using voxel-based deconvolution has also been proposed [12]. This method is based on the MLEM reconstruction algorithm and requires the determination and the optimization of eight parameters, most of them depending on the image quality or the PET scanner properties. Despite the fact that standard deconvolution algorithms are easy to implement, they suffer from significant noise propagation as a result of the additive (or multiplicative) approach in the regularization component. These noise properties force the application of deconvolution algorithms to specific ROIs, where activity concentration values are calculated and subsequently improved iteratively.

The aim of this study was to compare the performance of two alternative deconvolution methodologies (Lucy-Richardson and Van-Cittert) in combination with the introduction of dedicated wavelet-based denoising algorithms. The main objective of this work was to reduce the noise introduced by the deconvolution process and subsequently improve quantitative accuracy for voxel-wise PVE correction in whole-body imaging. The different algorithms considered were tested on various datasets, including simulated and phantom studies and clinical whole-body FDG PET images.

## Materials and methods

### Deconvolution

The general framework in deconvolution is based on the following model:

$$\mathrm{I}(\vec{\mathrm{r}}) = \mathrm{O}(\vec{\mathrm{r}}) \otimes \mathrm{PSF}(\vec{\mathrm{r}}) + \mathrm{N}(\vec{\mathrm{r}}) \tag{1}$$

where I is the observed image, O is the real object, PSF is the degrading PSF, N is an additive noise and $\otimes$ the convolution operator. Teo et al. in [11] used the Van Cittert algorithm [13] according to which it is possible to iteratively retrieve the actual object O from the observed data I (the emission tomography image in our case) and PSF, by writing:

$$O^{(n+1)}\left(\vec{r}\right) = O^{(n)}\left(\vec{r}\right) + \alpha\left(I\left(\vec{r}\right) - \mathrm{PSF}\left(\vec{r}\right) \otimes O^{(n)}\left(\vec{r}\right)\right) \tag{2}$$

where $\alpha$ is a converging parameter generally taken as 1. In this equation, the quantity $\mathrm{I}(\vec{\mathrm{r}}) - \mathrm{PSF}(\vec{\mathrm{r}}) \otimes \mathrm{O}^{(\mathrm{n})}(\vec{\mathrm{r}})$ is called the residual because it is obtained by subtraction and it converges towards noise. Considering this notation, the Van Cittert algorithm may be rewritten as:

$$\mathrm{O}^{(\mathrm{n}+1)}(\vec{\mathrm{r}}) = \mathrm{O}^{(\mathrm{n})}(\vec{\mathrm{r}}) + \alpha \mathrm{Res}^{(\mathrm{n})}(\vec{\mathrm{r}}) \tag{3}$$

For this reason the regularization step is additive.

There are a number of classical deconvolution algorithms, each one with specific characteristics. In the Lucy-Richardson deconvolution approach (see, for example, references [14] and [15]), the regularization step is multiplicative instead of additive:

$$O^{n+1}(\vec{\mathrm{r}}) = O^{n}(\vec{\mathrm{r}})\left[\frac{I(\vec{\mathrm{r}})}{I^{n}(\vec{\mathrm{r}})} \otimes PSF(-\vec{\mathrm{r}})\right] \quad \text{with } I^{n}(\vec{\mathrm{r}}) = PSF(\vec{\mathrm{r}}) \otimes O^{n}(\vec{\mathrm{r}}) \tag{4}$$

Considering the previous notation concerning the residual, Eq. 4 can be rewritten as:

$$O^{n+1}(\vec{\mathrm{r}}) = O^{n}(\vec{r})\left[\frac{I^{n}(\vec{r}) + \mathrm{Re}\,s^{n}(\vec{r})}{I^{n}(\vec{r})} \otimes PSF(-\vec{r})\right] \tag{5}$$

Compared with the Van-Cittert approach, this multiplicative process has the advantage of limiting noise propagation. For this reason, our investigation included the implementation of this technique for PVE correction in PET imaging in comparison to the Van-Cittert methodology, which was also included in this study.

## Wavelet-based denoising

In order to reduce noise introduced by the deconvolution process and thus facilitate the use of deconvolution based on the image rather than ROI for PVE correction in PET, we propose in this work the integration of wavelet-based denoising in the deconvolution process.

When considering image processing from a general point of view, wavelet denoising is one of the most powerful denoising methods and is the theme of numerous publications. The procedure consists of three steps, and the main hypothesis is that the observed image $I(\vec{r}) = S(\vec{r}) + N(\vec{r})$ contains the true signal S with noise N. These three steps are described by the following equations:

$$Y = W(I) \tag{6}$$

$$J = T(Y, t) \tag{7}$$

$$S_{est} = W^{-1}(J) \tag{8}$$

where W and $W^{-1}$ are the direct and inverse wavelet transform operators, T(t) is the denoising operator depending on the threshold t. In terms of the implementation of the denoising operator T and/or the threshold t selection different algorithms may be considered. In this work one of the most robust algorithms, the BayesShrink [16] described below, was used.

### *BayesShrink*

Given threshold t for any data w, the rule:

$$T(w, t) = sgn(w)\max(0, |w| - t) \tag{9}$$

defines the simplest version of thresholding, known as nonlinear soft thresholding. The operator T nulls all values of w for which $|w| \leq t$ and shrink towards the origin by an amount t all values of w for which $|w| > t$. Considering the context of use in the present study w stands for the wavelet values that we wish to threshold. For this reason, the soft thresholding rule in the wavelet domain is often referred to as wavelet shrinkage denoising. Most work concerning wavelet-based denoising is due to Donoho [17], with one of the primary developments being a method called Visushrink [18]. In this latter, a general threshold, now usually referred to as universal threshold, is defined as $Tu = \sqrt{2\sigma^2 \log N}$ with N being the sample size and σ the noise expressed as the standard deviation of the wavelet transform values. Although in most common situations it is impossible to measure σ from the corrupted (noisy) image, it is possible to estimate it from the first subband of the wavelet transform as:

$$\sigma = \frac{Median(|w_i|)}{0.6745},\ w_i \in \text{first sub} - \text{band} \tag{10}$$

Without challenging the soft-thresholding strategy, alternative threshold value selections have been also proposed. Assuming a generalized gaussian distribution of the wavelet coefficients, Chang et al. proposed a method based on the Bayes theory, referred to as BayesShrink [16]. This approach performs soft-thresholding with a data-driven, subband-dependent threshold $T_B = \sigma^2/\sigma_X$ with $\sigma^2$ denoting the noise variance estimated using the median operator already described, and $\sigma_X = \sqrt{\max(\sigma_w^2 - \sigma^2, 0)}$ where $\sigma_w^2 = \frac{1}{n^2}\sum_{i=1}^{n^2} w_i^2$ and $n^2$ is the size of the considered subband. The main interest in this robust method is because of the fact that the filtering threshold is adapted to different levels of resolution where noise properties can be very different. Actually, because of the limited PSF, the noise in emission images is not white but coloured and therefore noise variance changes across subbands.

## The wavelet transform

Wavelets have already been used in different domains and different applications of emission tomography [9, 19, 20]. In this section we consider the choice of the wavelet transform provided that some requirements are met concerning the specific application of interest. First, the kind of images we are concerned with may be large and thus require a substantial amount of computer memory. Furthermore, the overall processing time should remain reasonable since use in the clinical domain is the aim. At last, a crucial point is the fact that we need to reconstruct the image from its altered (filtered) wavelet coefficients. For this reason, the so-called Decimated Wavelet Transforms (for example the pyramidal algorithm of Mallat) are not well adapted for our application due to the loss of the translation-invariance property [21, 22]. The elimination of the decimation process leads to wavelet transforms of size exactly equal to the original image, as in the dyadic "à trous" algorithm. This latter approach to wavelet transform is called Undecimated Wavelet Transform (UWT). At each iteration, which decreases spatial resolution while preserving image size, three images are generated presenting edges and details in horizontal, vertical and diagonal directions respectively. Equivalence between the UWT and Mallat's

algorithm has been previously demonstrated by Shensa [23]. In some applications, as in the one we are dealing with, the direction may appear of reduced practical interest. By choosing adequate filters it is possible however to obtain a single isotropic wavelet image instead of three directional ones allowing a gain in both memory requirement and computing time. This approach is called Isotropic Undecimated Wavelet Transform (IUWT) and was chosen for our denoising problem. From a theoretical point of view, its similarity with Mallat's algorithm has been recently proved by Starck et al. [24]. A 2-D wavelet decomposition was chosen for the sake of simplicity.

*Algorithm*

The IUWT of the image I, performed using the filter f, leads to the set $W = \{w_1, w_2, \ldots, w_i, I_J\}$, where $w_i$ are the wavelet coefficients at scale i and $I_J$ is a smoothed version of I, that is I at scale J. The passage from one level of resolution, or scale, to the next which is reduced by half (dyadic progression) is performed as follows:

$$I_{i+1}(x,y) = f \otimes I_i = \sum_{u,v} f(u,v) I_i\left(x + 2^i u, y + 2^i v\right) \qquad (11)$$

The corresponding wavelet coefficients, on which actual denoising is performed, are hence obtained by a simple pixel-to-pixel subtraction:

$$w_{i+1}(x,y) = I_i(x,y) - I_{i+1}(x,y) \qquad (12)$$

Finally, the original image I is reconstructed without any loss by adding all wavelet images to the final smoothed $I_J$:

$$I(x,y) = I_0(x,y) = I_J(x,y) + \sum_{j=1}^{J} w_j(x,y) \qquad (13)$$

PVE correction by integrating wavelet-based denoising into iterative deconvolution

In the previous sections we have introduced the different tools to be employed for the implementation of the proposed voxel-wise PVE correction in such a way that the complete methodology can now be described in detail. The algorithm consists of only two nested steps.

In the first step the image we wish to correct for PVEs undergoes deconvolution using either the Van-Cittert or the Lucy-Richardson algorithm. The denoising phase is performed concomitantly by applying wavelet-based thresholding to the residual *Res* (Eqs. 3 and 5 for the Van-Cittert and the Lucy-Richardson algorithms, respectively) at each iteration of the deconvolution process. Thus, before addition or multiplication, *Res* undergoes IUWT (section The wavelet transform) and the obtained wavelet images are denoised according to the BayesShrink method (section BayesShrink). In the present study, the IUWT was performed using the bicubic-spline filter and up to J=3, which means that three wavelet images (or subbands) were obtained. Each of these was denoised and the obtained altered wavelet values were used to reconstruct a noise-attenuated residual *Res*. One of the limitations of wavelet-based thresholding is the 2-D aspect. To circumvent this limitation the denoising process was applied on the three planes of the space (axial, coronal, sagittal) and the final thresholded *Res* image was obtained by averaging the three sets of data on a voxel-by-voxel basis. For comparison purposes, deconvolution without denoising was also performed using both the Van-Cittert and Lucy-Richardson methodologies.

Deconvolution algorithms assume that the PSF is known and stationary across the field of view (FOV), or at least inside the whole image to be analysed. This is not the case in emission tomography where the PSF can vary in the three directions of the FOV. Although an accurate measure of the PSF is possible by acquiring images of line sources at different positions throughout a system's FOV, one can only use a single PSF in the deconvolution process. This could potentially induce errors when considering the accuracy of the PVE correction in different regions of a given image. In the previous work by Teo et al. [11], it was shown that a 1-mm error in the PSF would lead to a negligible impact on standard uptake values measured inside homogeneous areas. This study concerned the Van-Cittert algorithm only and no denoising technique was incorporated into the deconvolution process. We further investigated this potential issue in our study considering an image simulated with a 6-mm PSF and containing a 15-mm diameter sphere. The sphere-to-background ratio (SBR) was modelled as 8:1 (before degrading by Poisson noise and a 6-mm PSF). SBR was finally measured after deconvolution using the different algorithms considered in this study with different PSF values, ranging from 4 mm to 8 mm, in 1-mm increments.

All programs were implemented using C on a Pentium 4, single processor, 2 Gb memory personal computer.

Test images

Different images were used to test and compare both the Van-Cittert and Lucy-Richardson algorithms, with and without wavelet denoising. Simulated and phantom data were used to assess the behaviour of the algorithms and to quantify their properties, considering perfectly known object intensity and sizes.

The aim of the first image dataset was to appreciate the global behaviour of the techniques and was manually designed from a real heterogeneous liver tumour isolated

from a clinical FDG scan. The tumour was extracted from the original image and then manually segmented into three classes (necrosis, active tissue, background). Activity values were attributed to each class based on the relative activity levels in the original image and Poisson noise was added with mean value equal to each considered class value. The obtained image was then blurred by a 6.5 mm × 6.5 mm × 7 mm 3D- Gaussian kernel, giving the final synthetic tumour PET image. The same kernel size was used in the deconvolution algorithms.

The second dataset comprised real acquisitions from a FDG-filled cylinder phantom (IEC) containing six spheres with diameters ranging from 10 mm to 37 mm. Acquisitions were carried-out in the list-mode format using a Philips GEMINI GXL PET/CT scanner (Philips Medical Systems, Cleveland, OH). The SBR was set to 8:1 by introducing 7.4 kBq.cm$^{-3}$ and 59.2 kBq.cm$^{-3}$ in the background and in the spheres, respectively. Three different statistical qualities were obtained by reconstructing 1-min, 2-min and 5-min list-mode time frames using the 3-D RAMLA algorithm (voxel sizes of 2 cm×2 cm×2 cm). These phantom images were also used to assess the impact of the accuracy of the PSF size to the overall performance of the deconvolution algorithms considered (see also section PVE correction by integrating wavelet-based denoising into iterative deconvolution).

The third set of images comprised clinical FDG whole-body images of 13 patients undergoing oncological staging. Images were acquired on a Philips GEMINI GXL PET/CT scanner (seven patients) at an average of 54 min after injection of an average of 380 MBq with an acquisition of 3 min per bed position, and on a GE Discovery LS (GE Healthcare, UK) scanner (six patients) after an injection of an average of 366 MBq and an acquisition time per bed position of 5 min. Lesions were localized in the lungs and/or in the abdomen. Images of the patients acquired with the Philips and GE systems were reconstructed using optimized parameters for the RAMLA 3-D [25] and OSEM (two iterations, 28 subsets, 4.3×4.3×4.25-mm voxels) respectively.

### Data collection and analysis

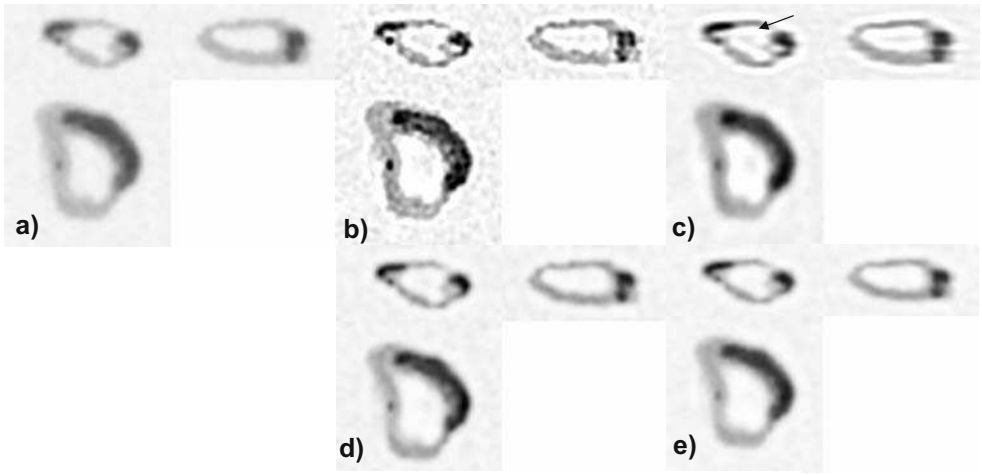
Intensity was computed as the mean signal inside ROIs. For the simulated tumour and the cylindrical phantom, the size of the ROIs was known exactly and their actual boundaries were used to calculate mean intensities. The level of noise inside these ROIs was calculated as the standard deviation of the signal. For this purpose we used smaller ROIs in order to estimate the noise in areas as homogeneous as possible. The aim was to eliminate the regions close to the boundaries where on the one hand PVEs are known to produce large intensity gradients and where on the other hand voxels contain a mixture of tissues due to discrete sampling (tissue fraction effect).

Considering the clinical data we first evaluated the intensity inside large and visually uniform areas including the lungs, the liver and other soft tissues. The selection of these ROIs was made assuming that PVEs were either null or negligible inside them. In total 45 ROIs were manually drawn. The mean intensity and the noise were thus calculated inside areas of identical sizes, contrary to the lesions where the same approach as for the simulated/phantom data was used. Nevertheless, the exact size of the lesions was obviously unknown. As a consequence they were manually segmented by an expert in the field of clinical nuclear medicine imaging.

## Results

The different PVE correction methods implemented in this study, i.e. Van-Cittert with and without BayesShrink denoising, and Lucy-Richardson with and without BayesShrink denoising, are referred to from here onwards as VC_B, VC, LR_B and LR, respectively. Images in Fig. 1 show the synthetic tumour simulated from an actual liver tumour, before and after using the different deconvolution algorithms. One of the things to notice is the very noisy aspect of the image generated by the Van-Cittert deconvolution alone (Fig. 1b). When incorporating wavelet-based denoising in this algorithm, noise is drastically reduced but artefacts present in the originally deconvolved images (see Fig. 1b) persist, manifested by underestimated activity regions appearing around high-intensity tissues (see arrow, Fig. 1c, and plot profile, Fig. 1f). On the contrary, the Lucy-Richardson algorithm provided artefact-free images, with a far less significant noise even without denoising (Fig. 1d, e). The quantitative results, which are given in Fig. 2, show that both deconvolution methods added noise to the corrected image. Nevertheless, the use of the Van-Cittert algorithm alone always led to the largest amount of noise while the combination of Lucy-Richardson and wavelet-based denoising provided less noisy images (in certain circumstances approaching the noise levels of the original image; for example, tissue 1 in Fig. 2a). Considering the accuracy of the PVE correction (i.e. the ability to retrieve the original intensity in a given tissue), the percentage of intensity loss is shown in Fig. 2b. Tissue 1 (expected intensity 20,000) and tissue 2 (expected intensity 50,000) intensity losses due to PVEs are 10% and 15%, respectively, while the application of the LR_B protocol led to 3.8% and −0.6%, respectively, which are the best results among all the PVE correction methods under investigation.

The same PVE correction processes were applied to the IEC cylinder phantom images containing spheres of



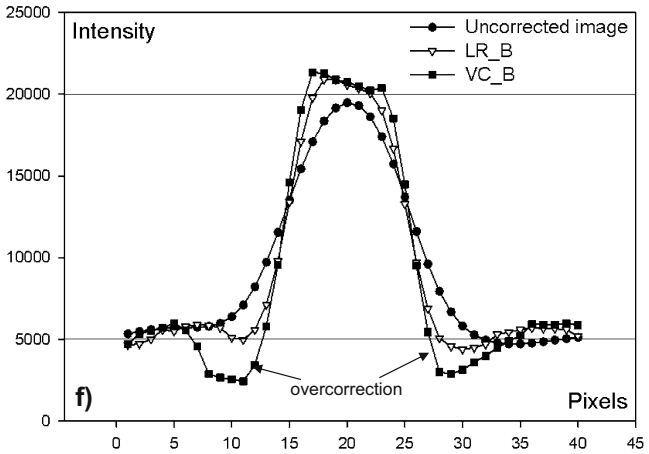




**Fig. 1** Synthetic tumour simulated from the manual segmentation of a real liver tumour. **a** The uncorrected image shows two classes of different active tissues surrounding a necrotic area. **b–e** Images after application of the VC method (**b**), VC_B method (**c**), LR method (**d**) and LR_B method (**e**). **f** Plot profile across the area shown by the arrow in **c**. Note the significant overcorrection induced by the VC_B method

different diameters (expected SBR of 8:1). Qualitatively similar remarks as for the synthetic tumour can be made (Fig. 3). Van-Cittert deconvolution alone generated elevated noise levels both in the background and in the spheres (Fig. 3b). Adding wavelet-based denoising significantly attenuated these effects (Fig. 3c), although LR_B method provided the best visual results in this example (Fig. 3e). The SNRs (calculated as mean/SD) in the background area

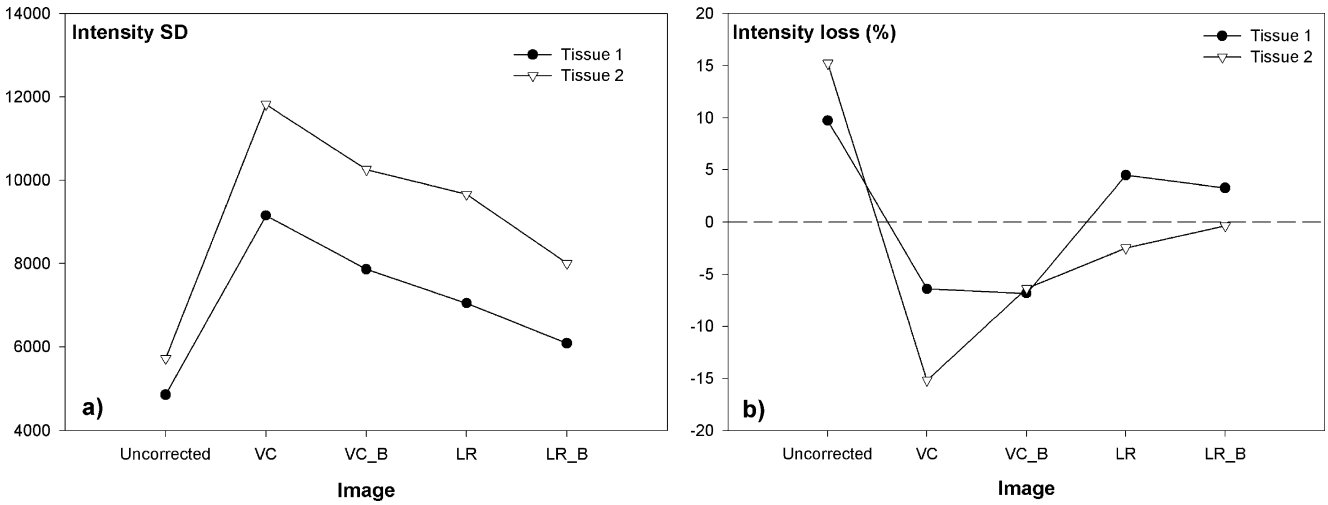



**Fig. 2** Quantitative measures in the tissues of the synthetic tumour. **a** Noise amount expressed as the standard deviation of the intensity in each tissue. **b** Accuracy of the intensity recovery expressed as the error percentage between the expected intensity (20,000 and 50,000 for tissue 1 and tissue 2, respectively) and the measured mean intensity in each tissue

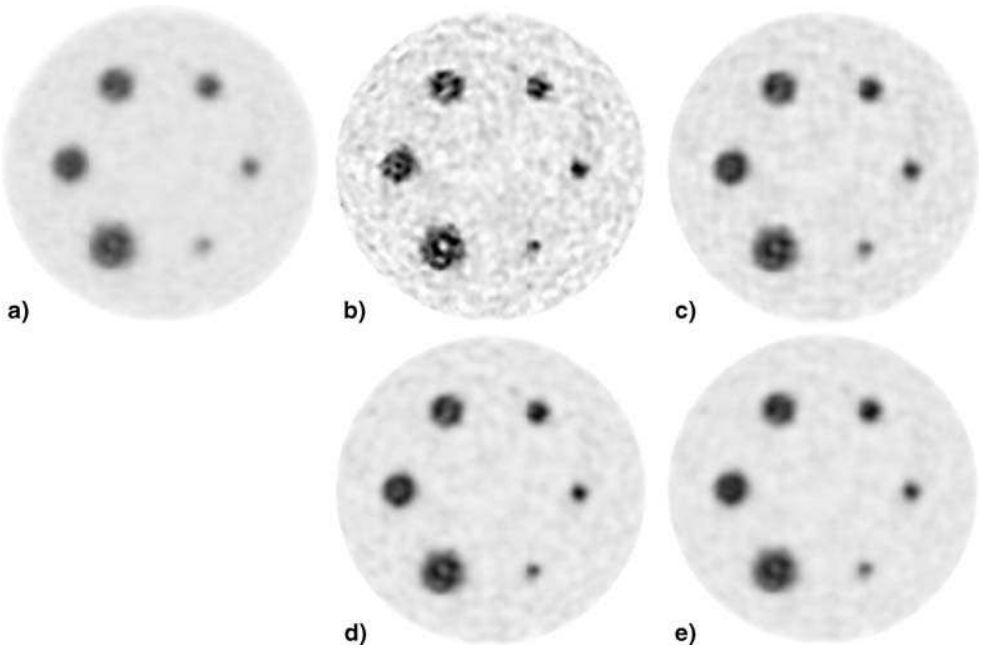



**Fig. 3** IEC phantom containing spheres. **a** Uncorrected image, transverse view. **b–e** Transverse views after application of the VC method (**b**), VC_B method (**c**), LR method (**d**) and LR_B method (**e**)

of the cylinder for the uncorrected image, the VC image, the VC_B image, the LR image and the LR_B image were: 9.54, 3.23, 3.96, 5.51 and 6.67, respectively, for the 1-min time frame acquisition; 12.08, 4.87, 5.91, 7.61 and 8.63, respectively, for the 2-min time frame acquisition; and 16.41, 6.90, 8.13, 10.82 and 12.34, respectively, for the 5-min time frame acquisition. These values support the visual impression that the use of the Van-Cittert deconvolution alone leads to large noise levels, significantly compromising any potential PVE correction for the derivation of improved images.

SBRs and percentage noise increase are shown on Fig. 4. In this particular set of phantom data the percentage of SBR recovery was found to be comparable whatever the algorithm used, either with or without denoising. The intensity recovery difference in the spheres considering the different algorithms under investigation was at a maximum between the images corrected with the VC and the LR_B algorithms (difference between 5% and 20% increasing with decreasing lesion size). These results were generally uniform whatever the level of statistical noise in the images as a result of evaluating different time acquisitions (1, 2 and 5 min; Fig. 4a, Fig. 4b and Fig. 4c, respectively). We also observed that no method was able to retrieve the original 8:1 SBR, with the absolute 8:1 ratio being more closely approached for large than for small spheres. The percentage of SBR increase was significant, ranging from 10.1±1.9% (largest sphere) to 40.9±2.0% (smallest sphere). The impact of the denoising step on the actual PVE correction was found to be negligible. More specifically we measured a global SBR decrease of only 1.2±0.4 when passing from a deconvolution method without denoising to the same method with denoising.

The results concerning the level of noise in the PVE-corrected images are shown in Figs. 4d–f. Several observations can be made. Firstly, the statistical quality of the original (uncorrected) images depending on the acquisition time (from 1 min to 5 min) appears to have no correlation with the level of noise introduced by deconvolution. In contrast, the size of the object of interest appeared to be more correlated with noise introduction since a greater noise increase was seen for the three largest spheres in comparison to the three smallest ones. Another important result concerns the actual noise levels in PVE-corrected images. The VC algorithm led to an increase in noise of between 82% and 362% (see Fig. 4e). The LR method always performed better, even if significant noise levels were also introduced in some cases. The use of wavelet-based denoising in the VC deconvolution led to a reduction in the amount of noise (absolute percentage decrease of −12% to −94%, mean decrease −37±24%).

It is important to note that the Lucy-Richardson deconvolution performed better even without denoising and in all the cases included in the study. LR-corrected images showed 121±30% less noise than VC_B-corrected images. In three cases (Fig. 4e, f) LR correction led to the generation of images with even less noise than the original (uncorrected) images. The incorporation of wavelet-based denoising into the LR iterative process led to the lowest levels of noise among all cases, whatever the statistics of the initial image and the sphere diameter. The average noise percentage difference between the LR and LR_B methods, considering all the different phantom configurations under evaluation (such as acquisition time and sphere size; Fig. 4), was 26%. In summary, (1) VC and LR led to the same level of PVE correction, (2) LR outperformed VC in

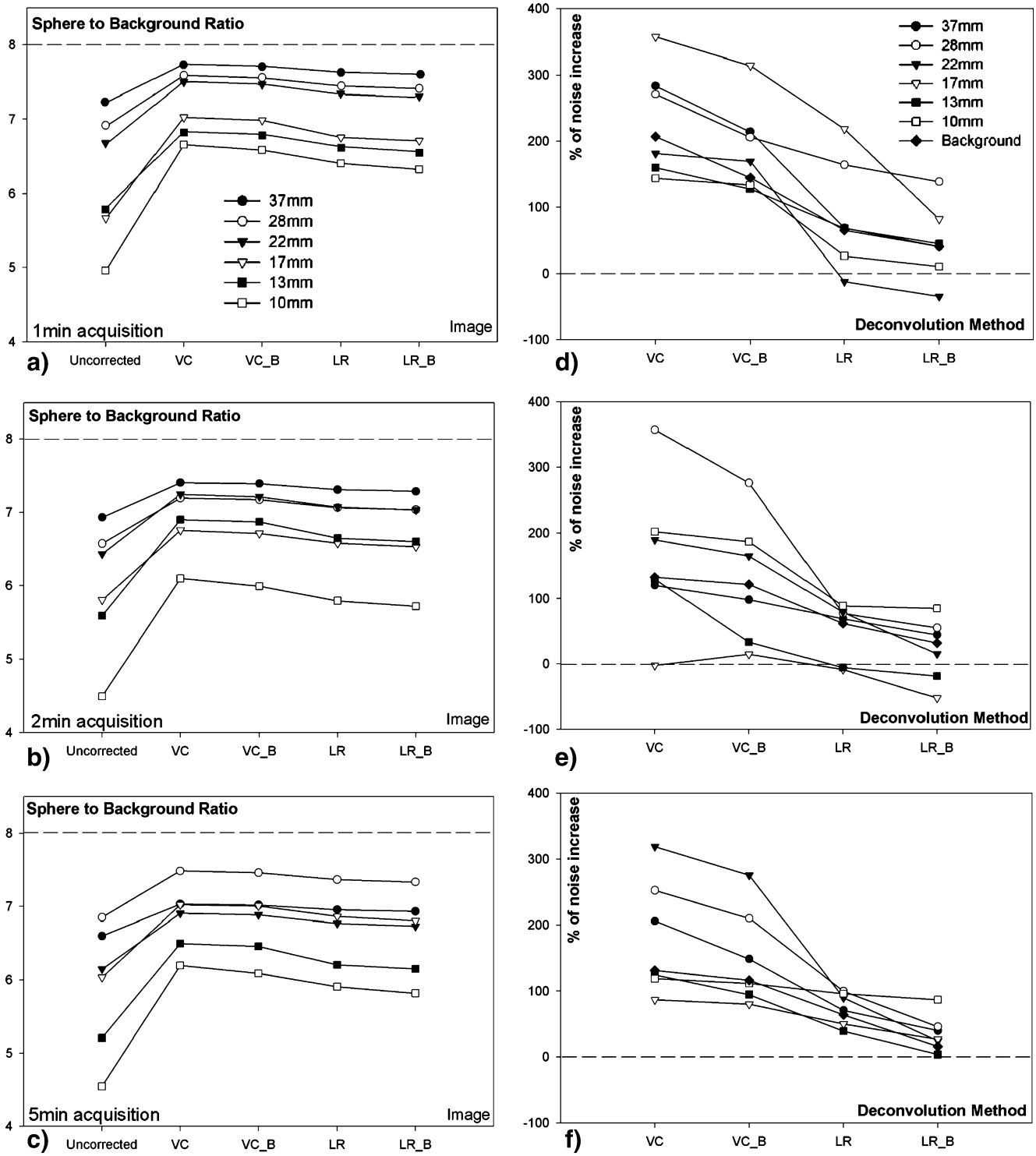



**Fig. 4** SBR (**a**–**c**) and noise increase percentage (**d**–**f**) in the six spheres of the IEC phantom. **a**, **d** 1-min time frame acquisition image; **b**, **e** 2-min time frame acquisition image; **c**, **f** 5-min time frame acquisition image

terms of noise propagation, and (3) wavelet-based denoising was effective in reducing overall noise without altering the quantitative intensity recovery.

Figure 5 shows the results obtained using the FDG whole-body images from one of the 13 patients. Considering the ROIs in large and visually uniform tissues (for example lung and liver) the application of the LR_B method to the whole-body images led to an average change of 0.6±2.8% in the mean intensity of large and uniform areas. The results obtained with the other methods were 7.2

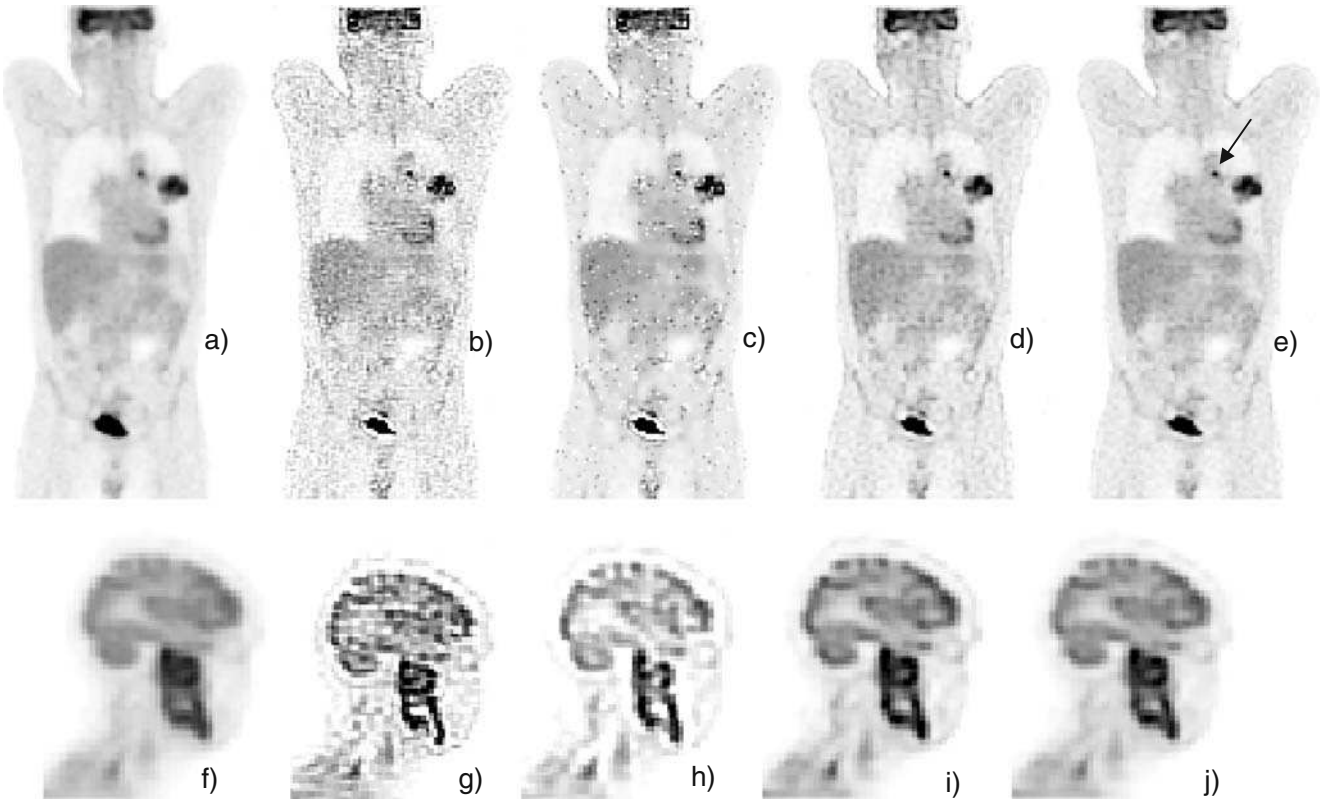



**Fig. 5** Clinical images: examples of PVE correction using the proposed approaches. **a–e** Patient 1, coronal view (mediastinal and pulmonary lesions).**f–j** Patient 2, sagittal view (head and neck lesion). **a**, **f** Uncorrected images. **b–e**, **g–j** Images after application of the VC method (**b**, **g**), VC_B method (**c**, **h**), LR method (**d**, **i**) and LR_B method (**e**, **j**). For patient 1, using LR_B, the mean intensity in the small lesion (**e** *arrow*) is increased by 23% and at the same time the image quality is preserved

±21.4% for the VC method, 6.7±16.6% for VC_B method and 0.8±2.6% for the LR method. A similar trend was seen for the noise levels (expressed as standard deviation) in the same set of ROIs. Applying the Van-Cittert algorithm to the whole-body images led to an increase by as much as three times in noise levels relative to the original (uncorrected) image (+195±112%). Although wavelet denoising reduced the introduced noise by a factor of nearly 2, there was still a significant increase (+113±97%) relative to the original (uncorrected) images. On the other hand, relative to the uncorrected images, the amounts of noise amplification as a result of the use of the LR and the LR_B algorithms were +29.3±14.1% and +19.1±13.4% respectively. Similar conclusions in terms of both intensity recovery and noise propagation can be drawn for the set of ROIs in the patient lesions (22 ROIs in total, mean size 38±26 $cm^3$). The mean increases in intensity for these ROIs were 59.1±25.5%, 39.5±17.2%, 37.4±13.8% and 27.5±8.8% for the VC, VC_B, LR and LR_B methods, respectively. Concerning noise the observed values were 260.8±101.1%, 124.0±64.4%, 100.9±42.1% and 49.2±14.4%, respectively.

Knowledge of the PSF is the only requirement for the algorithms presented in this study. It is thus necessary to evaluate the influence of the PSF on the quantitative results obtained. The data obtained by varying the PSF size are shown in Fig. 6. Firstly, the effect of the PSF value on the SBR accuracy appear linear since underestimation of the PSF leads to underestimation of the SBR while overesti-

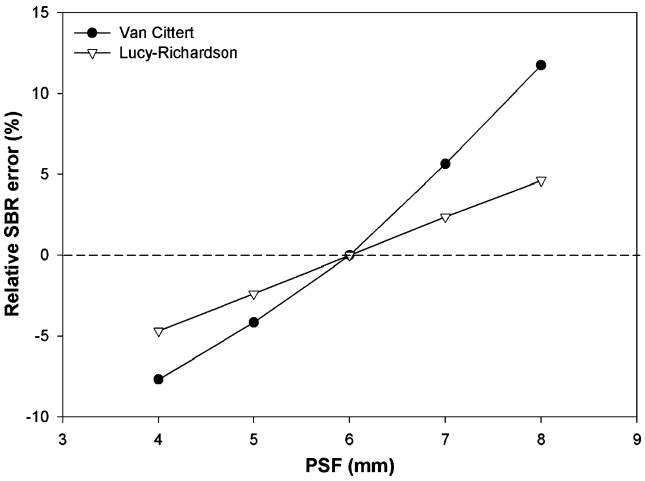



**Fig. 6** Relative SBR expressed as percentage errors, obtained with different PSF values. The expected SBR value corresponds to that obtained with the actual PSF (6 mm). For example, the use of a 4 mm PSF instead of the actual value in the Lucy-Richardson algorithm leads to a 5% SBR underestimation

mation of the PSF leads to overestimation of the SBR. A significant point also not illustrated here is the fact that incorporating wavelet-based denoising in either of the two deconvolution algorithms considered did not modify the PSF dependence compared with deconvolution without denoising. Thus there was no effect of denoising on the PSF value or on SBR estimation. The second point is that an error of up to 2 mm in the PSF led to a maximum 5% error in the SBR estimation when using the Lucy-Richardson deconvolution. In comparison, an error of 1 mm in the PSF led to a 5% error in the SBR estimation when using the Van Cittert algorithm, which is in accordance with the study of Teo et al. [11].

The results of the same study in each sphere of the IEC physical phantom are shown in Fig. 7. Although similar results were seen, a supplementary point can be underlined. When investigating small lesions (less than 15 mm in diameter), one should pay attention to the estimation of the PSF because even with the L-R algorithm a ±1 mm error in the PSF estimation may lead to a 5–7% error in the SBR estimation, although under the same imaging conditions the error is nearly 20% with the VC algorithm.

## Discussion

PVEs are difficult to overcome and remain a major problem in emission tomography, in particular when investigating small lesions. This is typically the case in FDG whole-body studies, which are today an essential tool in the management of patients with a tumour. Most PVE correction algorithms are, however, restricted to brain imaging because of the mandatory use of anatomical information provided by MRI. This morphological requirement is difficult to transpose to whole-body imaging since small structures are not easy to superimpose even with dedicated PET/CT scanners. This is mainly because of both internal (physiological) and external patient motion during and between acquisitions. For these reasons, classical PVE correction methods cannot be applied in whole-body imaging with a sufficient degree of reliability.

Recently, iterative deconvolution-based algorithms have been proposed for tumour imaging [10–12]. In the study by Teo et al. [11], only the Van Cittert algorithm was tested and showed the introduction of high levels of noise, which led the authors to propose the application of the algorithm only to specific ROIs (such as a well-delineated tumour). In the study by Tohka and Reilhac [10], two different deconvolution algorithms were tested, but only in the specific case of raclopride brain PET imaging. In the study by Kirov et al. [12], the method did not require anatomical images and noise was not added during the process, and the method required few parameters to be manually defined by the user. The objectives of our study were to assess two fully automatic and iterative deconvolution algorithms for whole-body PET imaging, and to introduce modifications to limit noise propagation without significantly affecting the PVE correction aspects.

The findings of the present study regarding the use of iterative deconvolution algorithms in oncology PET imaging allow us to draw several conclusions. First, automatic PVE attenuation can be achieved at a voxel level with a fairly limited noise increase. The intensity recovery in relatively small areas is indeed significant enough to accept the moderate noise addition inherent to iterative deconvolution. However, on the other hand our results suggest that the choice of algorithm is crucial. In our investigation, the Van-Cittert approach led to larger increases in signal recovery (by a factor of up to 40%) in comparison to the Lucy-Richardson approach. On the other hand, an increase

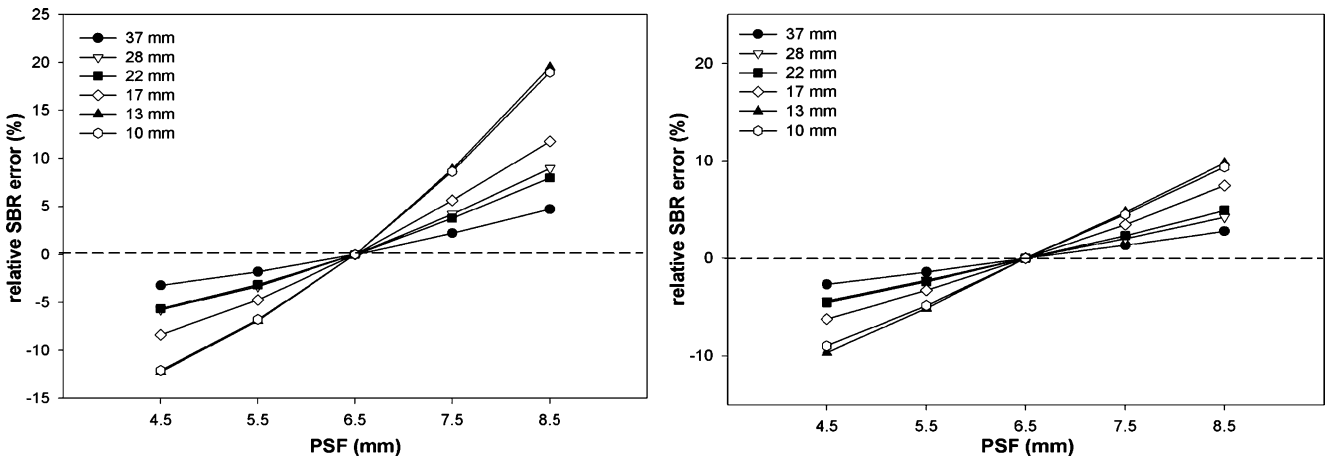



**Fig. 7** Relative SBR obtained in the IEC phantom for different PSF values. For a given sphere, the expected SBR value corresponds to the one obtained with the actual PSF (6.5 mm). **a** Van-Cittert algorithm; **b** Lucy-Richardson algorithm

of >250% in the image noise levels was also measured with the use of the Van-Cittert algorithm in comparison to the uncorrected images, while the noise introduced with the use of the Lucy-Richardson was smaller by a factor of three relative to the Van-Cittert algorithm. The levels of noise associated with the Van-Cittert algorithm led to images which were visually difficult to interpret, while artefacts were also observed around high-intensity regions. Finally, changes in intensity were also observed in areas of the body where uniform activity distributions are usually observed (for example, the normal lung and liver).

All these factors suggest that the Van-Cittert algorithm strongly and systematically alters the actual intensity in regions irrespective of their size. This bias affecting the voxel values in both a random and a systematic fashion clearly indicates that the use of the Van-Cittert algorithm for quantitative purposes should be avoided, a point that is also in accordance with the qualitative assessment of images. Comparatively, the Lucy-Richardson deconvolution provided not only good qualitative performance but also satisfactory quantitative improvement as well. Unlike the Van-Cittert method, the Lucy-Richardson method led to more consistent results in both qualitative and quantitative assessments irrespective of the imaging conditions (noise level, lesion size). These specific conclusions apply to both simulated and clinical images. Finally, it is worth noting that the Lucy-Richardson algorithm is as easy to implement as the Van-Cittert method (the only difference is in the iteration step which is multiplicative in the Lucy-Richardson method and additive in the Van-Cittert method).

One of the major developments presented here concerns the introduction of a denoising step during regularization of the deconvolution process, in an attempt to minimize noise propagation by the algorithms. The denoising procedure was implemented in the wavelet domain in an attempt to minimize the potential parallel loss of pertinent signal leading to an associated loss in resolution. Although a 2-D wavelet decomposition was employed, the denoising operation was performed on the three planes (axial, coronal, sagittal) with the final thresholded residual image obtained by averaging the three sets of data on a voxel-by-voxel basis. The BayesShrink method was chosen for the selection of the denoising operator and threshold values. This approach has the benefit of adapting the threshold to the noise properties observed at different levels of resolution [16]. However it should be noted that depending on the image reconstruction algorithm used, the noise within each level of resolution (wavelet subbands) considered may not be homogeneous (particularly in the case of iterative reconstruction algorithms). This heterogeneity in noise will not be accounted for by the adaptive threshold of the denoising operator proposed here, although it is expected to have a limited impact on the results obtained. A possible way to account for such noise heterogeneity in the different subbands is the use of more recent denoising approaches [26]. Our results clearly demonstrate that the inclusion of denoising as proposed here has the potential to significantly reduce (by a factor of >100%) the level of noise introduced by the deconvolution process.

As a final point in relation to this study, one can say that the Lucy-Richardson algorithm was less sensitive to the PSF value than the Van Cittert algorithm, and an error of up to 2 mm remains tolerable regarding the associated changes in SBR. Thus the Lucy–Richardson algorithm is expected to be less sensitive to changes in the PSF that can be observed throughout a system's FOV.

## Conclusion

We have shown that deconvolution algorithms for the correction of PVEs must be used with care. The application of a denoising step following deconvolution appears mandatory in order to be able to perform reliable voxel-wise quantitative analyses. In particular, the use of the Van Cittert deconvolution alone was shown to propagate noise to an amount that prevented reliable quantitative evaluation. A significant bias was also observed in some cases leading to over-correction, while the process always led to visual degradation of the images. The use of the Lucy-Richardson algorithm together with wavelet-based denoising appeared to give more robust and consistent results in both accuracy and precision, while preserving the visual aspect of the images.

## References


1. Kato H, Shimosegawa E, Oku N, Kitagawa K, Kishima H, Saitoh Y, et al. MRI-based correction for partial volume effect improves detectability of intractable epileptogenic foci on 123I-iomazenil brain SPECT images. J Nucl Med 2008;49:383–9.
2. Kalpouzos G, Chetelat G, Baron JC, Landeau B, Mevel K, Godeau C, et al. Voxel-based mapping of brain gray matter volume and glucose metabolism profiles in normal aging. Neurobiol Aging 2009;30:112–24.
3. Mevel K, Desgranges B, Baron JC, Landeau B, De la Sayette V, Viader F, et al. Detecting hippocampal hypometabolism in Mild Cognitive Impairment using automatic voxel-based approaches. Neuroimage 2007;37(1):18–25.
4. Samuraki M, Matsunari I, Chen WP, Yajima K, Yanase D, Fujikawa A, et al. Partial volume effect-corrected FDG PET and grey matter volume loss in patients with mild Alzheimer's disease. Eur J Nucl Med Mol Imaging 2007;34(10):1658–69.
5. Rousset OG, Ma Y, Evans AC. Correction for partial volume effects in PET: principle and validation. J Nucl Med 1998;39 (5):904–11.
6. Rousset OG, Collins DL, Rahmin A, Wong DF. Design and implementation of an automated partial volume correction in PET: application to dopamine receptor quantification in the normal human striatum. J Nucl Med 2008;49:1097–106.

7. Basu S, Alavi A. Feasibility of automated partial-volume correction of SUVs in current PET/CT scanners: can manufacturers provide integrated, ready-to-use software. J Nucl Med 2008;49:1031–2.
8. Soret M, Bacharach SL, Buvat I. Partial-volume effect in PET tumor imaging. J Nucl Med 2007;48(6):932–45.
9. Boussion N, Hatt M, Lamare F, Bizais Y, Turzo A, Cheze-Le Rest C, et al. A multiresolution image based approach for correction of partial volume effects in emission tomography. Phys Med Biol 2006;51(7):1857–76.
10. Tohka J, Reilhac A. Deconvolution-based partial volume correction in raclopride-PET and Monte Carlo comparison to MR-based method. Neuroimage 2008;39:1570–84.
11. Teo BK, Seo Y, Bacharach SL, Carrasquillo JA, Libutti SK, Shukla H, et al. Partial-volume correction in PET: validation of an iterative postreconstruction method with phantom and patient data. J Nucl Med 2007;48(5):802–10.
12. Kirov AS, Piao JZ, Schmidtlein CR. Partial volume effect correction in PET using regularized iterative deconvolution with variance control based on local topology. Phys Med Biol 2008;53:2577–91.
13. Van Cittert PH. Zum einfluss der spaltbreite auf die intensita tsverteilung in spektrallinien. Z Physik 1931;69:298.
14. Lucy LB. An iteration technique for the rectification of observed distributions. Astron J 1974;79:745–54.
15. Richardson WH. Bayesian-based Iterative Method of Image Restoration. J Opt Soc Am 1972;62(1):55–9.
16. Chang SG, Yu B, Vetterli M. Adaptive wavelet thresholding for image denoising and compression. IEEE Trans Image Process 2000;9(9):1532–46.
17. Donoho DL. De-noising by soft-thresholding. IEEE Trans Inf Theory 1995;41(3):613–27.
18. Donoho DL, Johnstone IM. Ideal spatial adaptation via wavelet shrinkage. Biometrika 1994;81:425–55.
19. Turkheimer FE, Aston JA, Asselin MC, Hinz R. Multi-resolution Bayesian regression in PET dynamic studies using wavelets. Neuroimage 2006;32(1):111–21.
20. Kalifa J, Laine A, Esser PD. Regularization in tomographic reconstruction using thresholding estimators. IEEE Trans Image Proc 2003;22(3):351–9.
21. Starck JL, Murtagh F, Bijaoui A. Image processing and data analysis: the multiscale approach. Cambridge: Cambridge University Press; 1998.
22. Mallat S. A theory for multiresolution signal decomposition: the wavelet representation. IEEE Trans Pattern Anal Mach Intell 1989;11:674–93.
23. Shensa MJ. Discrete wavelet transform: wedding the à trous and Mallat algorithms. IEEE Trans Signal Proc 1992;40 (10):2464–82.
24. Starck JL, Fadili J, Murtagh F. The undecimated wavelet decomposition and its reconstruction. IEEE Trans Image Proc 2007;16(2):297–309.
25. Visvikis D, Turzo A, Gouret S, Damien P, Lamare F, Bizais Y, et al. Characterisation of SUV accuracy in FDG PET using 3D RAMLA and the Philips Allegro PET scanner. J Nucl Med 2004;45:103P.
26. Ramani S, Blu T, Unser M. Monte-Carlo SURE: a black-box optimization of regularization parameters for general denoising algorithms. IEEE Trans Image Proc 2008;17:1540–54.

CONTRIBUTED PAPER

# Incorporating Patient-Specific Variability in the Simulation of Realistic Whole-Body $^{18}$F-FDG Distributions for Oncology Applications

*Computed data, that describes the anatomy and breathing-motion of individual cancer patients, is used to increase the realism of computer models that represent the patients bodies.*

By Amandine Le Maitre, William Paul Segars, *Member IEEE*, Simon Marache, Anthonin Reilhac, Mathieu Hatt, *Member IEEE*, Sandrine Tomei, Carole Lartizien, and Dimitris Visvikis, *Senior Member IEEE*

**ABSTRACT** | The purpose of the work described in this paper was the development of a framework for the creation of a realistic positron emission tomography (PET) simulated database incorporating patient-specific variability. The ground truth used was therefore based on clinical PET/computed tomography (CT) data of oncology patients. In the first step, the NURBS-based cardiac-torso phantom was adapted to the patient's CT acquisitions to reproduce their specific anatomy while the corresponding PET acquisitions were used to derive the activity distribution of each organ of interest. Secondly, realistic tumor shapes with homogeneous or heterogeneous activity distributions were modeled based on segmentation of the PET tumor volume and incorporated in the patient-specific models obtained at the first step. Lastly, patient-specific respiratory motion was also modeled. The derived patient-specific models were subsequently combined with the PET SORTEO Monte Carlo simulation tool for the simulation of the whole-body PET acquisition process. The accuracy of the simulated datasets was assessed in comparison to the original clinical patient images. In addition, a couple of applications for such simulated images were also demonstrated. Future work will focus on the creation of a comprehensive database of simulated raw data and reconstructed whole-body images, facilitating the rigorous evaluation of image-processing algorithms in PET for oncology applications.

**KEYWORDS** | Database; Monte Carlo simulation; positron emission tomography (PET); SORTEO

Manuscript received February 27, 2009; revised June 8, 2009 and July 1, 2009. Current version published November 18, 2009. This work was supported in part by the Brittany region, France, under Grant QUANTEP.

**A. Le Maitre**, **M. Hatt**, and **D. Visvikis** are with LaTIM, INSERM, 29609 Brest, France (e-mail: amandine.lemaitre@etudiant.univ-brest.fr; hatt@univ-brest.fr; Visvikis.Dimitris@univ-brest.fr).
**W. P. Segars** is with the Radiology Department, Duke University, Durham, NC 27706 USA (e-mail: paul.segars@duke.edu; wsegars@jhmi.edu).
**S. Marache**, **S. Tomei**, and **C. Lartizien** is with CREATIS-LRMN, 69621 Lyon, France (e-mail: simon.marache@creatis.insa-lyon.fr; sandrine.tomei@creatis.insa-lyon.fr; carole.lartizien@creatis.insa-lyon.fr).
**A. Reilhac** is with Biospective Inc., Montreal, PQ, Canada (e-mail: anthonin@biospective.com).

Digital Object Identifier: 10.1109/JPROC.2009.2027925

## I. INTRODUCTION

Positron emission tomography (PET) functional imaging using $^{18}$F-FDG is widely considered as the state of the art in diagnosis for a number of oncology applications [1]. In addition, current interest within the clinical context concentrates on further extending the impact of PET imaging in other applications, such as patient follow-up during treatment and assessment of tumor response to therapy

(either chemotherapy or radiotherapy) [2]. In order to optimize the quantitative use of PET in clinical practice, data- and image-processing methods are also a field of intense interest and development. The evaluation of such methods often relies on the use of simulated data and images since these offer control of the ground truth. Monte Carlo (MC) simulations are widely used for PET simulation since they take into account all the random processes involved in PET imaging, from the emission of the positron to the detection of the photons by the detectors. In general, MC simulations play an important role for different clinical applications, notably in the domains of medical imaging and radiotherapy [3]–[6]. In the field of medical imaging, simulations are primarily used for the design and optimization of new and existing imaging devices at the level of detectors and associated electronics systems [3], [7]. The second area of interest lies in the use of simulations for the assessment of acquisition and subsequent processing protocols involving the production of raw datasets and corresponding reconstructed images. Numerous simulation packages are now available to create such datasets, like GATE [8], SimSET [9], or SORTEO [10]. GATE offers additional precision considering physics modelling but is also much more computationally demanding. SimSET, on the other hand, is faster despite the use of a precise physical modelling but is limited in terms of the geometry of the detectors it can model. PET-SORTEO has the advantage of being fast with a precise geometry modelling but is less precise than GATE as far as the physics modelling is concerned. A complete review of available simulation tools for emission tomography can be found in [11].

In the area of emission tomography, simulated raw datasets can be used for the optimization of reconstruction algorithms and associated correction strategies such as, for example, those for scattered detected events [12], [13]. Based on these simulated datasets, reconstructed PET images are often used to test various image-processing algorithms, examples of which include partial volume effects correction [14], denoising [15], automated detection [16], and segmentation [17]. Using MC simulated datasets is the most reliable approach to validate and assess the performance of such algorithms since they include an accurate modelling of all physical processes associated with the acquisition process. In addition, the advantage of using simulated data over clinical data is that it allows control of the ground truth that is usually not available in the case of patient studies. However, phantoms offer limited shapes and realism, and the use of more realistic and clinically relevant datasets, in terms of geometry and activity distribution, should improve the evaluation of methodologies destined to be used in clinical practice. Within this context a few simulated PET databases already exist. A first one was developed by Reilhac *et al.* using PET-SORTEO [10], [18]. It contains brain images modelling tracer kinetics based on the use of different radiopharmaceuticals. A second one was developed by Castiglioni *et al.* [19] containing several images of simulated brain and thorax parts of the Zubal phantom [20] as well as cylindrical phantoms generated using various simulation tools. Aristophanous *et al.* [21] have created a single whole-body phantom simulation based on the use of the Zubal phantom in combination with the SimSET code, incorporating nonuniform lesions in the lungs. However, no patient variability in terms of anatomical and functional activity distribution or physiological motion (such as respiration) was considered. More recently, a third database of FDG whole-body images was developed by Tomei *et al.* [22] using PET-SORTEO. This database is based on the use of the Zubal phantom including spherical lesions of various sizes with different contrast using a model of lesion extent based on the clinical description of lymphoma patients.

The aim of this work is to improve the realism of simulated whole-body PET images through the incorporation of both anatomical and functional uptake variability considering different disease models. Within this context, the simulated images include variability from patient to patient, taking into consideration the anatomy of each patient as well as specific $^{18}$F-FDG distribution in organs and lesions. In addition, the proposed database includes the effects of respiratory motion through the simulation of four-dimensional (4-D) whole-body PET images based on the respiratory signals of individual patients.

## II. MATERIALS AND METHODS

The simulation process consists of two major steps. As a first step, the model of the patient's anatomy is created, followed by the simulation of the PET acquisition based on the developed patient's model. This simulation requires the model of a scanner geometry. The individual steps are explained in the following sections.

### A. Clinical Data

The datasets to be simulated are based on whole-body clinical images. The objective is to simulate images which are as realistic as possible, modelling the anatomical variability and corresponding tumors from the original clinical images. The clinical images used up to date were acquired on a PET/CT Philips GEMINI system (Philips Medical Systems, Cleveland, OH). The total injected activity was between $2.78 \times 10^5$ kBq and $5.10 \times 10^5$ kBq and the images were acquired 1 h postinjection. The scan time for each bed position is 2 min. These clinical images were reconstructed with the iterative reconstruction algorithm RAMLA [23] used in Philips GEMINI scanners and standard clinical protocol parameters previously optimized (two iterations, a Gaussian postfiltering with 3-D full width at half-maximum (FWHM) of 5 mm and a relaxation parameter of 0.05) [24]. CT scan acquisitions were carried out using 120 kV and 100 mA for tube settings, resulting in CTDI from 600 to 800 mGy. The voxel size of the CT

Table 1 Summary of the Ten Clinical Scans Used for the Simulation

| | Sexe | Age | Weight (kg) | Injected Activity (kBq) | Cancer type | Tumour Description | |
|---|---|---|---|---|---|---|---|
| | | | | | | Localisation | Number |
| Patient1 | Male | 79 | 77 | $3.82x10^5$ | Lung cancer | Left lung | 1 |
| Patient2 | Male | 47 | 62 | $3.09x10^5$ | Lymphoma | Spleen | 2 |
| Patient3 | Male | 55 | 71 | $3.53x10^5$ | Lymphoma | ORL | 1 |
| Patient4 | Male | 64 | 78 | $3.88x10^5$ | Lymphoma | Abdomen | 6 |
| Patient5 | Female | 55 | 66 | $3.42x10^5$ | Lymphoma | Left Lung | 1 |
| Patient6 | Female | 77 | 60 | $3.37x10^5$ | Lymphoma | Mediastinum | 5 |
| Patient7 | Female | 66 | 75 | $3.70x10^5$ | Lymphoma | Right Lung/Mediastinum | 2 |
| Patient8 | Male | 53 | 65 | $2.98x10^5$ | Lung cancer | Right Lung | 1 |
| Patient9 | Female | 80 | 55 | $2.78x10^5$ | Lung cancer | Left Lung | 1 |
| Patient10 | Male | 56 | 94 | $5{,}10x10^5$ | Lung cancer | Left Lung | 1 |

images was $1.17 \times 1.17 \times 5$ mm$^3$ with dimensions of $512 \times 512 \times N$ ($N$ slices depending on the patient's axial coverage). The voxel size of the PET images was $4 \times 4 \times 4$ mm$^3$ with dimensions of $144 \times 144 \times M$ ($M$ also depending on the patient's axial coverage). We selected ten patients (six men and four women) with different shapes of tumors at different positions in the body and with homogeneous and heterogeneous activity distributions. The different patient data with corresponding injected activity and tumor characteristics are summarized in Table 1. The CT images were used for the anatomy modelling and the PET images for the activity distribution modelling.

The model obtained by the following procedure is a voxel-based description of the patient's anatomy. Each organ is designated by one label. These labels have to be associated with both an activity concentration and an attenuation coefficient in order for the simulation tool to reproduce the process of radioactive disintegration, particle interaction, and photon detection.

### B. Modelling of the Anatomy

*1) Organ Shape:* The anatomical and activity distribution models were constructed using the nonuniform rational basis splines (NURBS)-based cardiac-torso (NCAT) phantom [25] as a basis. This model achieves a good balance between flexibility and realism thanks to the use of NURBS surfaces for the organ shape modelling. In order to take into account the interpatient anatomical variability, we used an interactive software application [26], which allows the modification of the NCAT phantom's anatomy. We thus modified the NCAT model based on specific patient anatomy using the CT clinical images acquired during the corresponding patient PET/CT studies as a guide. Two-dimensional slices of the NCAT phantom were overlaid with corresponding slices of the CT clinical images. The organ shapes were modified by changing the position of the control points associated to the NURBS surfaces of each organ. Within this process, the first step involved the use of simple transformations like rotations, translations, or scaling applied to one or several organs via their respective control points. This first step allows for a global adaptation of the phantom to the anatomy of a given patient. A finer adjustment was subsequently applied by a displacement of individual control points to more accurately reshape the organs of interest. The majority of the different organ shapes were adapted to those of the specific patients, although some differences still remained for some complex organs of the abdomen. We found, in particular, that the intestines are difficult to match with those of specific patient acquisitions because of large differences between the modelled shape in the NCAT phantom and the real shape in combination with large interpatient variability in both shape and form. A potential improvement for the matching of organs, such as the intestines, between the model and specific patient will be the use of deformable models, which have not been explored in this work.

*2) Activity Distribution:* The clinical PET images were employed in order to accurately model the variable FDG distributions. A region of interest (ROI) analysis was performed for the different organs (liver, lungs, kidneys, etc.) of each patient in order to identify mean and standard deviation of activity concentrations per organ. In addition, these activity concentrations were compared to the standardized uptake values (SUVs) available in already published work of normal organ uptakes in FDG PET imaging [27] and used to assign the activities for the different organs in the NCAT patient specific emission maps used in the simulation.

Table 2 illustrates for four different organs (liver, lungs, stomach, and spleen) the comparison of the theoretical SUV (column 1) and the clinical SUV (column 2) obtained from the ROI analysis. For the ROI analysis, we considered 17 different organs: liver, lungs, heart wall, heart chamber, stomach, kidneys, spleen, spine bone, other bone, pelvis, bladder, intestines, rectum, ovaries/prostate, urethra, brain, and body (including muscle).

Table 2 Comparison of Theoretical SUV and Clinical SUV Measured on the Ten Clinical Datasets Used in the Study

| | Theoretical Mean SUV [27] | Clinical Mean SUV |
|---|---|---|
| Liver | 2.17 +/- 0.33 | 2.25 +/- 0.07 |
| Lungs | 0.48 +/- 0.10 | 0.41 +/- 0.01 |
| Stomach | 2.34 +/- 0.59 | 2.26 +/- 0.14 |
| Spleen | 1.65 +/- 0.29 | 1.93 +/- 0.07 |

*3) Attenuation Map:* The attenuation map was constructed using the same labelled phantom model as the one for the emission map. Each label of the phantom was associated with the corresponding attenuation coefficient at 511 keV. The different attenuation coefficients of the organs were approximated with those of seven structures: water (0.096033 $cm^{-1}$), bone (0.151108 $cm^{-1}$), lung (0.024667 $cm^{-1}$), brain (0.098530 $cm^{-1}$), fat (0.087718 $cm^{-1}$), air (0.000111 $cm^{-1}$), and muscle (0.098731 $cm^{-1}$).

## C. Tumor Modelling

The next step consisted in adding the tumor to the healthy patient model. The NCAT phantom allows the incorporation of homogeneous spherical tumors into the model. However, in order to simulate realistic patient-specific phantoms, we frequently need to incorporate modelling of complex tumor shapes and heterogeneous distributions. For this purpose, a new process for the incorporation of nonuniform shape and activity distribution lesions was developed in order to model lesion heterogeneity within the NCAT phantom.

The following steps were considered in this process. Initially, the tumors were manually segmented from the clinical PET images considering variable activity distributions and shapes. A 3-D reconstruction was subsequently derived using the segmented structures to obtain a mesh using AMIRA.[1] The organs in the NCAT phantom are modelled with NURBS. These types of surfaces allow more flexibility as well as the incorporation of respiratory motion. Since one of our aims is to include the simulation of patient-specific respiratory motion effects, the tumor surfaces were also modelled using NURBS for compatibility. The last step of tumor modelling is therefore the conversion of the mesh into a NURBS surface, for which *Rhinoceros* (CADLINK software) was used. More specifically, section curves were first extracted from the mesh of the tumor, and cubic NURBS surfaces were then fitted to these contours. The number of curves and the distance between them as well as the number of control points to use depend on the tumor shape. The more complex the shape is, the more curves and control points are needed for a realistic modelling. These parameters were therefore chosen in order to achieve the best match between each tumor mesh and its corresponding NURBS surface. Within two consecutive curves, a normal distribution is considered for the interpolation. The complete procedure is illustrated in Fig. 1.

In order to model tumor nonuniformities, as many NURBS surfaces as levels of activity identified within the tumors were created. During the creation of the voxelized phantom, each NURBS surface is associated with a specific level of activity. For the patients included to date in the simulated database under construction, only tumors with one or two levels of activity were modelled, since it was found to be sufficient for the adequate modelling of the nonuniform activity distributions encountered. However, the extension of the approach to include higher levels of nonuniformity in the modelled lesions is straightforward. In the work described in this paper, the decision regarding the number of activity levels to be included in the modelling of heterogeneity was based on a combination of visual and ROI assessments. ROIs were manually drawn within the visually distinguishable nonhomogeneous regions within the tumors, and the standard deviation with respect to the mean was measured. The standard deviations around the mean values within the ROIs were usually smaller than 10%. Additional activity regions would have hence led to insignificant nonuniform activity contrast differences with the main tumor.

Fig. 2 illustrates the result of the two first steps; it shows the coronal slices of a CT clinical image, the corresponding clinical PET slice, and the labelled phantom obtained by the procedure described above with the tumor incorporated in the right lung.

## D. Respiratory Motion

The default respiratory cycle of the NCAT phantom is a sinusoid [28]. In order to take into consideration the nonuniform nature of realistic respiratory signals, we replaced this cycle by patient specific respiratory signals. These were acquired during a 4-D PET/CT acquisition with an external sensor placed on the patient's thorax [29]. This resulting respiratory cycle is therefore nonregular in both phase and amplitude. It is characterized by a nonregular period and nonregular inspiration and expiration phases, contrary to the default NCAT respiratory cycle.

Each respiratory cycle was divided into $N$ bins, and one phantom was created for each of the $N$ instances of the respiratory cycle.

## E. Data Simulation

*1) Simulation Process:* The simulations of the images presented in this paper were carried out using the PET-SORTEO simulation tool that has been fully validated for the ECAT HR+ scanner geometry [10]. Six bed fields, each corresponding to the dimension of the scanner's axial field of view (FoV), were used during the simulation to achieve whole-body coverage (see Fig. 2), with the scanner

[1]http://www.amiravis.com/.

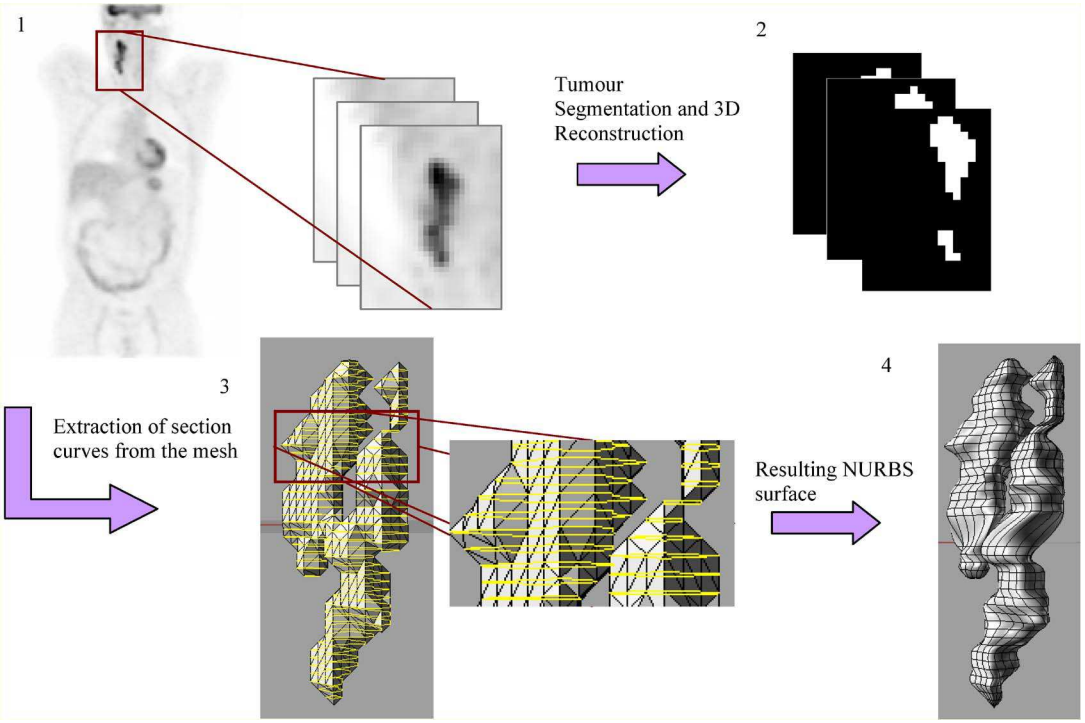



**Fig. 1.** ***Illustration of the tumor modelling process. The steps are: 1) tumor segmentation, 2) 3-D reconstruction of the tumor, 3) extraction of section curves from the mesh, and 4) reconstruction of a NURBS surface from the section curves.***

operating in 3-D. An overlap of 2.91 cm (corresponding to 12 reconstructed slices) was considered between two consecutive beds to compensate for the loss of sensitivity on both extremities of the axial FoV.

For the studies without the inclusion of respiratory motion, simulations were carried out approximating clinical acquisitions of 3–15 min per bed position in order to provide whole-body images of variable statistical quality. For the simulations with respiratory motion, the acquisition time per bed position used in static simulation ($T_{\mathrm{acqTot}}$) was divided by the number of respiratory cycles ($N_{\mathrm{bins}}$) and the number of bins per respiratory cycle ($N_{\mathrm{bins}}$) to obtain the acquisition time per bed position at one instance (bin) of the respiratory cycle ($T_{\mathrm{acqBin}}$)

$$T_{\mathrm{acqBin}} = \frac{T_{\mathrm{acqTot}}}{N_{\mathrm{cycle}} N_{\mathrm{bins}}}. \tag{1}$$

So the total acquisition time per bed position remains the same for simulations with or without the inclusion of

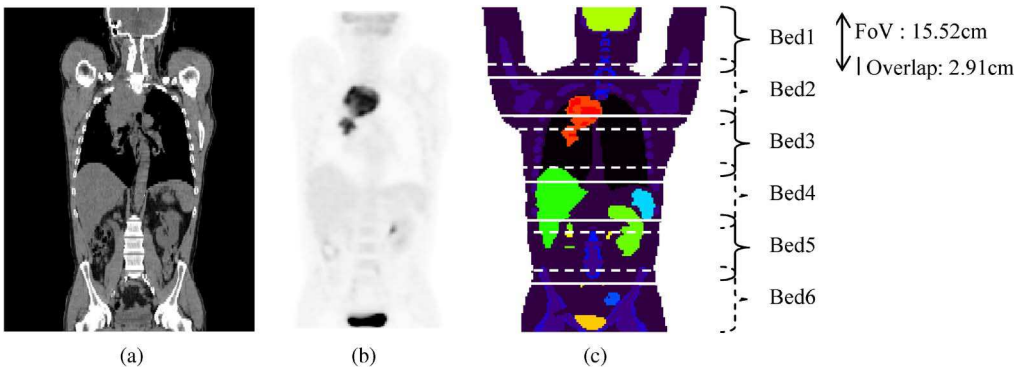



**Fig. 2.** ***Coronal slice of (a) a clinical CT image, (b) the corresponding clinical PET image, and (c) the corresponding slice of the adapted labelled NCAT phantom model. The six bed positions used for the simulation process are illustrated on the labelled phantom.***

respiratory motion. In order to obtain the gated simulated data, the sinograms at the same instance of the respiratory cycle were added before reconstruction. To obtain the respiratory average patient acquisition, all the sinograms were added before reconstruction.

Even considering the use of a simulation platform such as SORTEO, which is significantly faster than other alternatives, heavy load is put on computing resources by the simulation process. In order to generate data in a reasonable amount of time, we have been using the IN2P3 computing centre in Lyon.[2] Twenty processors were allocated for each bed of a whole-body simulation on Linux scientific 32-bit running systems. Such a configuration allowed us to generate a 14-min whole-body simulation without respiratory motion in about 90 h.

*2) Corrections and Reconstruction:* Before the reconstruction of the images, the standard corrections were applied to the simulated images: normalization [30], dead time, radioactive decay, scattered [31], and random corrections using a delayed coincidence window approach [18]. For the attenuation correction, two different approaches were used. In the first one, a transmission sinogram was employed. We simulated two acquisitions with a rotating $^{68}$Ge source, one with the patient model in place (3 min) and one in air [18]. The difference of the counts between the two acquisitions gives the attenuation coefficients, which can then be used for attenuation correction. An alternative way could be using CT-based attenuation correction. The simulation of CT images can be performed using the NCAT patient-specific attenuation maps created here and a dedicated CT simulator. This has been previously shown using standard NCAT maps and an analytical CT simulator [32]. However, since the simulation of patient-specific CT images was not the objective of the work described in this paper, an alternative approach, similar to the use of CT images for attenuation correction for the simulated emission images, was performed. The patient-specific NCAT attenuation maps (see Section II) were first filtered (Gaussian filter of $12 \times 12 \times 5$ mm$^3$) to match the resolution of the PET emission images [33] and subsequently projected forward [34] to derive the correction factors used for the attenuation correction.

The images were reconstructed with the iterative AW-OSEM algorithm, which is used clinically with the ECAT HR+ scanner [35]. The number of subsets (16) and iterations (four) used in the clinical setting was employed for the reconstruction of the simulated images. The final step was the use of a Gaussian isotropic filtering of 8 mm. Images were reconstructed using two different voxel sizes ($5 \times 5 \times 2.425$ mm$^3$ and $2.5 \times 2.5 \times 2.425$ mm$^3$), since this parameter represents an important factor influencing overall emission image quality and spatial sampling of the objects of interest (such as, for example, lesions in oncology applications). As such, it is important to be able to test the influence of voxel size on different image-processing algorithms, which is an example of the use of a database of simulated emission images introducing patient-specific variability, such as explained in this paper.

The different steps of the simulation process were detailed in this section. The following list of steps summarizes the procedure for the creation of one simulated image:

1) modelling of the patient-specific anatomy based on the CT clinical images;
2) description of the activity distribution in the main organs based on measurements carried out on the corresponding patient PET clinical images;
3) modelling of tumors (if necessary) by:
   a) segmenting the tumors and defining the various uptake levels within each tumor;
   b) transformation of the tumor mesh in a NURBS surface using Rhinoceros;
   c) incorporation of the tumor into the phantom;
4) definition of the scanner geometry within the chosen simulation environment;
5) simulation of the emission raw datasets (sinograms or list mode data) within the chosen simulation environment;
6) correction of the simulated raw datasets (sinograms or list mode data);
7) reconstruction of the PET simulated images.

## III. RESULTS AND APPLICATIONS

### A. The Whole-Body PET Simulated Database

To date we have generated ten different patients, eight of them corresponding to respiratory motion average PET/CT whole-body acquisitions and two of them corresponding to 4-D PET/CT acquisitions synchronized with respiratory motion. The respiratory motion average data were simulated with variable statistical qualities by varying the acquisition time as explained in Section II-D. A few examples of the simulated static PET whole-body images highlighting the wide range of patient variability in terms of anatomical and functional details are given in Fig. 3.

Two examples of tumors are shown in Fig. 4, where it can be seen that, in comparison with clinical data, the tumor shapes are qualitatively respected. In the clinical data, the voxel size is $4 \times 4 \times 4$ mm$^3$ whereas in the simulated datasets two different voxel sizes were used ($5 \times 5 \times 2.425$ mm$^3$ and $2.5 \times 2.5 \times 2.425$ mm$^3$). The difference in the contrast between the high uptake zone and the low uptake zone is about 2% in the second illustrated tumor; the first one was modelled with one level of activity only.

*1) Activity Distribution:* The injected dose in the cases of the clinical data on which we based the simulations of PET

[2]http://cc.in2p3.fr.

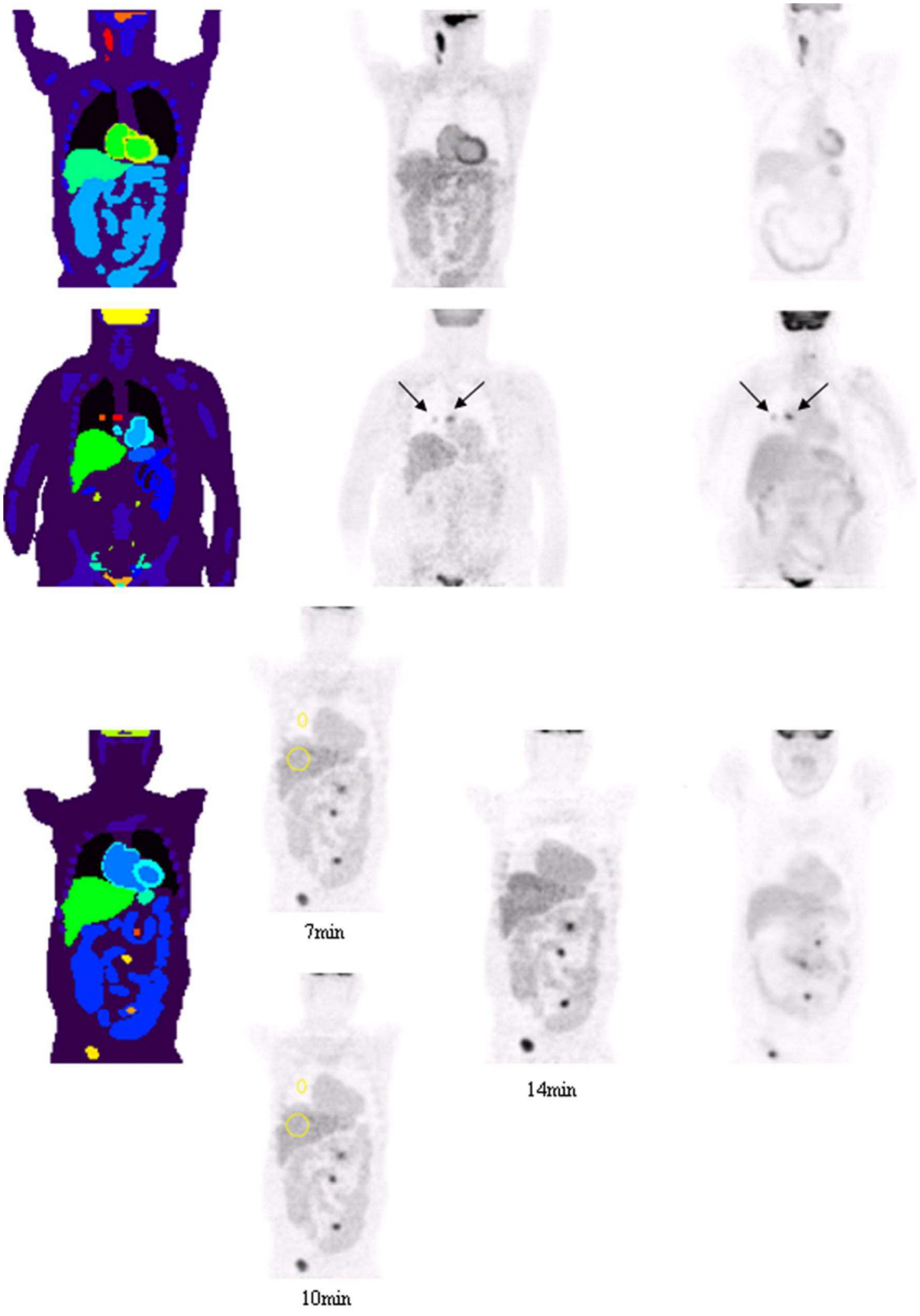
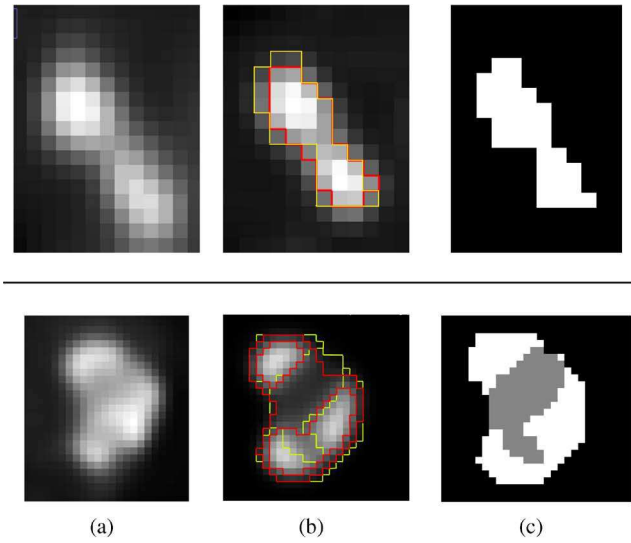



**Fig. 3.** ***Coronal slices from three patient whole-body simulated images: (a) the labelled NCAT phantom, (b) the simulated image, and (c) the clinical image.***




| | Simulated | Clinical |
|---|---|---|
| High activity level (kBq/mL) | 68.15 | 80.91 |
| Low activity level (kBq/mL) | 37.97 | 45.88 |
| Contrast | 0.56 | 0.57 |

**Fig. 4.** *Illustration of two different simulated tumors: (a) clinical image, (b) simulated image, and (c) labelled image. The table contains the differences in the corresponding activity concentrations and contrasts for the second tumor, which has been modeled using two activity levels. The results of the segmentation process are contoured in red, and the corresponding ground truth is shown in yellow.*

acquisitions ranged from $2.78 \times 10^5$ to $5.10 \times 10^5$ kBq. In the simulated data, the injected dose ranged from $3.10 \times 10^5$ to $4.68 \times 10^5$ kBq.

Fig. 5 contains the mean activity concentration differences for each organ of interest, derived using the ROI analysis, considering all eight static PET simulated images. A good agreement can be seen for most of the organs, with the highest mean activity differences in the kidneys, brain, and bladder (14%, 15%, and 23%, respectively), whereas for the other organs (liver, lungs, etc.), the mean difference is less than 9%.

Table 3 contains the standard deviation of mean activities within ROIs drawn inside organs of interest in order to illustrate the variability of the activity distribution from patient to patient. The patient variability of the simulated and the clinical data are of the same order of magnitude. The standard deviation in the heart wall is higher (about 5.3) than that of the other organs (mean standard deviation of about 1.3), which is consistent with the variable contrast seen in clinical PET images between the heart wall and the heart chamber.

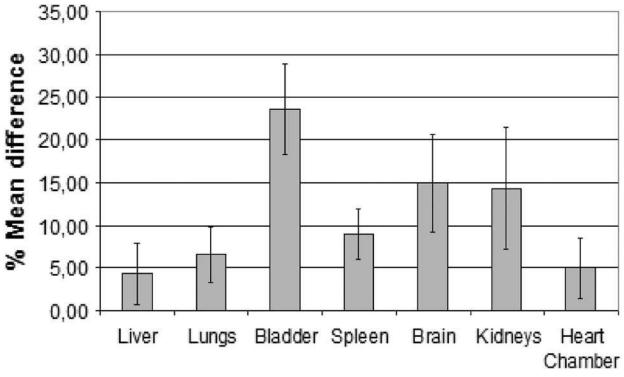



**Fig. 5.** *Mean percentage differences in the mean activity of the organs of interest between the simulated and the clinical images. The error bars denote the standard deviation in these mean differences for each organ.*

*2) Images of Different Statistical Quality:* For the third patient in Fig. 3, simulated images considering different acquisition times (see Section II-E) leading to different statistical quality, and hence various levels of noise, are illustrated. The mean activity concentration and standard deviation computed within the two ROIs drawn in the liver and the lungs were compared. The longer the acquisition time, the lower the standard deviation, with 24% and 23% decreases of the noise from 7 to 14 min for the liver and the lungs, respectively. On the other hand, the mean activity concentrations were largely unchanged (+/−1%).

*3) Dynamic Data:* Results on the gated acquisitions can be found in Fig. 6, which illustrates the resulting images at

Table 3 Comparison of the Mean and Standard Deviation of the Activities (in kBq/mL) Considering All of the Simulated Images for Organs of Interest in Both the Clinical and Simulated Datasets

| | Simulated | | Clinical | |
|---|---|---|---|---|
| | Mean | Std. Dev | Mean | Std. Dev |
| Liver | 12.33 | 1.08 | 11.97 | 1.60 |
| Lungs | 1.95 | 0.46 | 2.08 | 0.49 |
| Heart Chamber | 9.56 | 2.08 | 9.12 | 1.83 |
| Heart Wall | 13.35 | 5.24 | 13.56 | 5.31 |

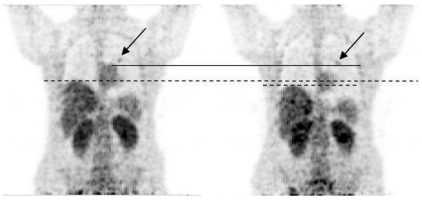


**Fig. 6.** *Coronal slices of the simulated whole-body scans for patient 5 at two instants (bin 1 and bin 6) of the respiratory cycle.*

two different times of the respiratory cycle for patient 5 (see Table 1). It corresponds to bins 1 and 6 of the respiratory cycle. These images were simulated with a spherical tumor with medium contrast (∼10) in the lung. The motion of the liver and the tumor in the lung are clearly visible in these two images. The top of the liver and the tumor are about 16 and 10 mm lower, respectively, in bin 6 with respect to bin 1. The estimated liver motion correlates well with the amplitude of the motion trace used in the simulation, which is expected considering the linear liver motion implemented in the NCAT phantom.

### B. Applications

In this section, we discuss two examples on the use of such a simulated database for the validation of PET image-processing methods. The first one is on the correction of partial volume effects and the second is on the automatic segmentation of tumor volumes.

*1) Partial Volume Effect Correction and Denoising:* Partial volume effect (PVE) is the consequence of the limited spatial resolution in PET. It affects both qualitative and quantitative accuracy of images as a result of the blurring and spill-in and spill-out effects. This is especially evident in tumor imaging, since the effect is major for structures of less than 1 cm in diameter, that is, twice the FWHM of the point spread function of the imaging device [36]. A large number of approaches have been proposed to correct for PVE [37]–[40], with some emerging methodologies concentrating on the production of improved image quality rather than concentrating on ROI analysis, which has been traditionally used for brain imaging applications [41]. In this application example, the method proposed by Boussion *et al.* [14] was applied to one of the illustrated simulated images (the second patient shown in Fig. 3). The simulated image contains two spherical tumors, one in the lung and one in the mediastinum, of 13 and 20 mm diameter, respectively. The PVE correction method is based on the use of PET data only and consists of the combination of a deconvolution and a denoising method in order to compensate for the noise amplification usually associated with iterative deconvolution methods. The Lucy Richardson deconvolution process is applied and the residual is denoised within each iteration using a wavelet-based denoising method. Such PVE correction leads to improved qualitative and quantitative analysis for oncology applications such as diagnosis and/or therapy assessment.

The original and corrected images are illustrated in Fig. 7, using a line profile demonstrating a 40% increase of the activity within the tumor in the mediastinum and a 28% increase for the tumor in the lung. A smaller increase relative to the ground truth is seen for the smaller lesion relative to the larger which agrees with the results reported on the original paper evaluating this PVE correction approach [14]. Table 4 shows the mean activity and associated standard deviation computed in the tumors and lungs in both simulated and corrected images. The tumor/lung contrast is

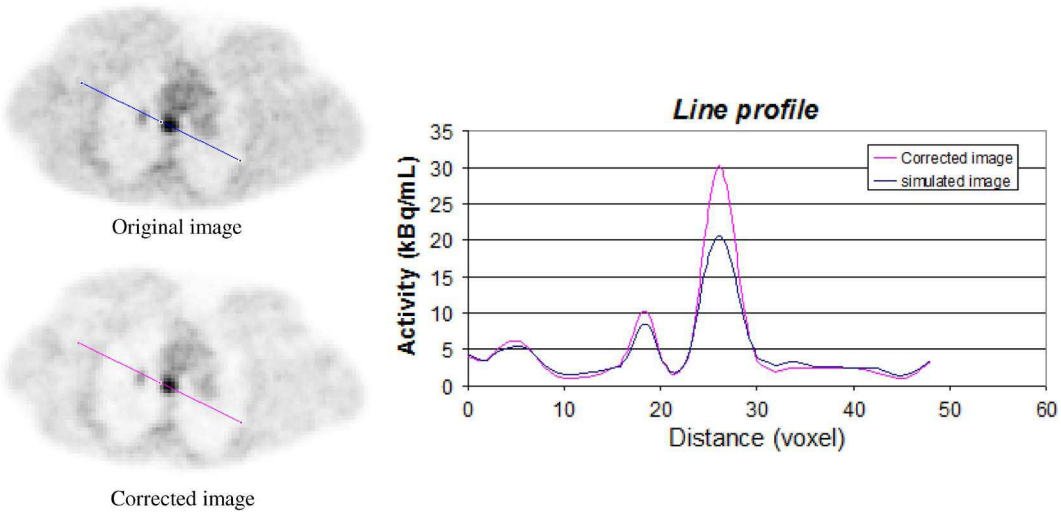



**Fig. 7.** *Result of PVE correction. The figure illustrates one transaxial slice of the simulated image (top) without and (bottom) with PVE correction. Two line profiles are plotted on the right.*

Table 4 Comparison With the Ground Truth of the Mean Activity (in kBq/mL) in the Lungs and the Tumors as Well as the Tumor/Lung Contrast in Simulated and Corrected Images

| | Simulated PET | | Corrected PET | | Ground Truth |
|---|---|---|---|---|---|
| | Mean | Std Dev | Mean | Std Dev | |
| Lung | 2.66 | 0.87 | 2.19 | 0.96 | 2.19 |
| Tumour1 | 9.1 | 2.19 | 11.6 | 2.81 | 18.44 |
| Contrast Tumour1/Lung | 3.4 | | 5.3 | | 8.4 |
| Tumour2 | 16.40 | 3.19 | 22.73 | 5.87 | 23.52 |
| Contrast Tumour2/Lung | 6.2 | | 10.4 | | 10.7 |

clearly improved after correction, with an increase from 3.4 to 5.3 in the case of tumor 1 and a similar improvement (from 6.2 to 10.4) for the second tumor, In addition, the derived mean values in the activity concentration of the lesions are closer to the simulated ground truth after correction, as shown in Table 4.

*2) Tumor Segmentation:* Another important clinical application in PET is tumor volume delineation for applications like patient follow-up and radiotherapy treatment planning, especially considering the growing interest of using hypoxia and proliferation tracers for improved biological target volume or dose painting and dose escalation applications [42]. The fuzzy locally adaptive Bayesian segmentation proposed by Hatt *et al.* [17] was applied here on the tumor illustrated in the tumor modelling process (see Fig. 1) and on the second heterogeneous tumor illustrated in Fig. 4 (quoted from here onwards as the second tumor).

The ground truth volume of the first and second tumor is 33.4 and 62.4 cm$^3$, respectively. The results of the segmentation are illustrated in Fig. 4 for the two tumors, showing one coronal slice for each of the two segmented tumors and the corresponding slice of the ground truth. The segmentation process results in an overall tumor volume of 35.2 cm$^3$ (5% error) and 61.8 cm$^3$ (1% error) for the first and second tumor, respectively.

## IV. DISCUSSION

The use of simulated datasets to evaluate new image-processing algorithms in PET is crucial. Indeed, it allows the control of the ground truth, which is not easily available when dealing with clinical datasets. In addition, Monte Carlo–based simulated images can be produced considering an accurate modelling of the overall detection process. In the field of emission tomography, a few simulated PET databases already exist, but the major concern remains the realism of these simulated images and the lack of variability between the different cases due to the use of nonflexible models. Within this context, the aim of the work described in this paper was to incorporate patient variability in a PET simulated database to improve the correspondence of the simulated images to the clinical cases. This variability concerns both the anatomy based on the clinical CT images and the $^{18}$F-FDG distribution based on the corresponding PET acquisition. The realism of our simulated images is also based on realistic tumor modelling's allowing variable tumor shapes and heterogeneity in the FDG uptake. Lastly, the possibility of generating respiratory motion in the simulated images was also included.

The patient-specific organ reshaping of the NCAT phantom, which forms the basic anthropomorphic model used in this paper, showed good results for the overall body shape and the organs of the thorax (lungs, liver, thoracic cage, etc.). A wide anatomical variability may be observed among the various images already simulated. Some problems, however, still remain at the abdominal region, particularly for the intestines as a result of the large interpatient variability at this level in combination with large differences between patient and model-based intestines.

Concerning the tumor modeling, the data simulated to date show good qualitative and quantitative correlation between the simulated and clinical data. In this paper, the choice of the number of activity levels needed to simulate heterogeneous tumors was made considering a combined visual and ROI analysis approach. However, this process could be improved using statistical measures of tumor heterogeneity [43].

For the images already simulated to date and included in the database, a good correlation between the simulated images and the clinical images was found in terms of activity distributions. The organs with the greatest differences between simulated and clinical FDG uptake are the kidneys, brain, and bladder. The differences measured in these specific organs can be largely attributed to the lack of realism in their modelling considering the version of the NCAT phantom used in this paper. The NCAT phantom, as its name indicates, was first developed for the thorax, with a main feature remaining the incorporation of respiratory and cardiac motions. Therefore, the main organs of interest are those located in the thorax. The brain, for instance, is only made of one simple structure, so the activity simulated in the brain is not as variable as the one existing in clinical images. For more realistic brain modelling, the Zubal phantom [20] is currently more precise. The kidneys and the bladder in the NCAT are also modelled with only one level of activity, whereas in reality, these are both heterogeneous organs in terms of FDG uptake.

The number of labelled regions available in the NCAT phantom may be a limiting factor for the definition of a more detailed activity map. For instance, in the version used in this paper, the muscles are not modelled. More specifically, there is only one structure named "body" around the different organs, and no distinction exists between the muscles and the fat. However, a later version of the NCAT phantom [44], developed for CT simulation, includes muscles and also more precise brain modelling. On the other hand, considering the availability of more structures within the NCAT phantom, there is no particular difficulty envisaged for the inclusion of additional labels in the workflow proposed in the work for the production of the simulated images.

The standard deviations in the different ROIs inside the organs could not be compared between the clinical and simulated images because the simulated images were not reconstructed with the same reconstruction algorithm (RAMLA for the clinical datasets and OSEM for the simulated images). In addition, the scanner geometry employed was not that corresponding to the system used in the acquisition of the clinical studies. Lastly, differences in terms of the correction algorithms used (attenuation, scatter, and random corrections) could also account for potential differences in the noise characteristics between the clinical and the simulated images. However, the purpose of this paper was to obtain realistic simulated FDG whole-body images rather than reproducing the exact clinical image quality of the Philips PET/CT system. Although this was not the objective of this paper, the simulation of different specific imaging devices can be more easily performed with alternative MC simulation codes than the SORTEO used here, which was specifically developed and validated for the Siemens ECAT HR+ scanner. For example, one such code is GATE (Geant4 Application for Tomographic Emission) [8], which facilitates an easier simulation of variable system architectures and associated readout and processing electronics [7]. The SORTEO MC simulation platform was chosen here because of its faster execution times in comparison to GATE at the beginning of the study. In any case, the approach described in this paper regarding the anthropomorphic model and its patient-specific adaptation is universal and applicable to any simulation platform.

The simulations including the respiratory motion effects as well as the use of a patient-specific respiratory cycle were validated as a first step and produced qualitatively acceptable images. The use of a patient-specific respiratory cycle improves the realism of the respiration modelling as it incorporates irregularities in the amplitude and the period of the motion. However, one should also note that the realism of the NCAT phantom (at least in the version used in this paper) in terms of respiratory motion at the diaphragm level is limited (models only a linear motion of the liver, the heart and the spleen). Lastly, the use of the actual version of the PET SORTEO simulation tool is not ideal for the simulation of respiratory motion since simulations including real-time physiological motion cannot be performed [45], and hence separate simulations have to be carried out for different instances in the respiratory cycle. In addition, the only raw data format available is sinograms, which can be limiting in comparison to list-mode format, which incorporates temporal information along the recorded coincidences for the investigation of dynamic processes and the application of associated correction algorithms [46]. Ongoing work will investigate the ability to generate list-mode data with the PET SORTEO simulation tool.

The execution times realized within the study described in this paper for the simulation of a whole-body PET image are reasonable within a research environment considering limited computing capabilities. Eventually, our objective will be to make available in a public database not only the simulated images but also the corresponding raw data files, which will facilitate the evaluation of image-reconstruction algorithms and associated corrections. In this paper, we have finally demonstrated on a couple of applications the usefulness of such simulated images in the validation and evaluation of postreconstruction processing algorithms, making clear use of the clinically realistic images and the availability of the ground truth.

## V. CONCLUSION

We have described a generic methodology for the development of a realistic simulated whole-body PET images database including patient-specific variability. The patient model was based on the NCAT anthropomorphic phantom adapted to patient-specific CT images. Ten simulated datasets containing tumors of various shapes and activity distributions were generated, with two of them including the simulation of respiratory gated PET frames, using patient-specific respiratory signals. The PET SORTEO simulation tool was used for the generation of the simulated datasets based on the derived patient-specific NCAT models.

Future work will include continuing the building of a comprehensive database of simulated static and 4-D FDG whole-body PET images from corresponding oncology patient acquisitions. The contents of the database will be complemented in the future by the raw simulated datasets as well as the use of alternative simulation platforms offering a larger spectrum of imaging system models and facilitating the simulation of 4-D processes. ■

## REFERENCES

[1] J. B. Bomanji, D. C. Costa, and P. J. Ell, "Clinical role of positron emission tomography in oncology," *Lancet Oncol.*, vol. 2, pp. 157–164, 2001.

[2] S. Ben-Haim and P. J. Ell, "FDG PET and PET/CT in the evaluation of cancer treatment response," *J. Nucl. Med.*, vol. 50, no. 1, pp. 88–99, 2009.

[3] H. Zaidi, "Relevance of accurate Monte Carlo modeling in nuclear medical imaging," *Med. Phys.*, vol. 26, no. 4, pp. 574–608, 1999.

[4] D. W. Rogers, "Fifty years of Monte Carlo simulations for medical physics," *Phys.*

*Med. Biol.*, vol. 51, no. 13, pp. R287–R301, 2006.

[5] E. Spezi, P. Downes, E. Radu, and R. Jarvis, "Monte Carlo simulation of an x-ray volume imaging cone beam CT unit," *Med. Phys.*, vol. 36, no. 1, pp. 127–136, 2009.

[6] D. Visvikis, M. Bardies, S. Chiavassa, C. Danford, A. Kirov, F. Lamare, L. Maigne, S. Staelens, and R. Taschereau, "Use of the GATE Monte Carlo package for dosimetry applications," *Nucl. Instrum. Methods*, vol. 569, pp. 335–340, 2006.

[7] F. Lamare, A. Turzo, Y. Bizais, C. Cheze-Le Rest, and D. Visvikis, "Validation of a Monte Carlo simulation of the Philips Allegro/Gemini PET systems using GATE," *Phys. Med. Biol.*, vol. 51, pp. 943–962, 2006.

[8] S. Jan, G. Santin, D. Strul *et al.*, "GATE: A simulation toolkit for PET and SPECT," *Phys. Med. Biol.*, vol. 49, no. 19, pp. 4543–4561, 2004.

[9] R. L. Harrison, S. D. Vannoy, D. R. Haynor, S. B. Gillipsie, M. S. Kaplan, and T. K. Lewellen, "Preliminary experience with the photon generator module of a public-domain simulation system for emission tomography," in *IEEE NSS-MIC Conf. Rec.*, San Francisco, CA, 1993, vol. 2, pp. 1154–1158.

[10] A. Reilhac, C. Lartizien, N. Costes, S. Sans, C. Comtat, R. N. Gunn, and A. C. Evans, "PET-SORTEO: A Monte Carlo-based simulator with high count rate capabilities," *IEEE Trans. Nucl. Sci.*, vol. 51, no. 1, pp. 46–52, 2004.

[11] I. Buvat and I. Castiglioni, "Monte Carlo simulations in SPET and PET," *Q. J. Nucl. Med.*, vol. 46, pp. 48–61, 2002.

[12] O. Barret, T. A. Carpenter, J. C. Clark, R. E. Ansorge, and T. D. Fryer, "Monte Carlo simulation and scatter correction of the GE advance PET scanner with SimSET and Geant4," *Phys. Med. Biol.*, vol. 50, no. 20, pp. 4823–4840, 2005.

[13] D. Lazaro, Z. El Bitar, V. Breton, D. Hill, and I. Buvat, "Fully 3D Monte Carlo reconstruction in SPECT: A feasibility study," *Phys. Med. Biol.*, vol. 50, pp. 3739–3754, 2005.

[14] N. Boussion, C. Cheze Le Rest, M. Hatt, and D. Visvikis, "Incorporation of wavelet based denoising in iterative deconvolution for partial volume correction in whole body PET imaging," *Eur. J. Nucl. Med. Mol. Imag.*, 2009, in press.

[15] F. Turkheimer, N. Boussion, A. Anderson, N. Pavese, P. Piccini, D. J. Brooks, and D. Visvikis, "PET image denoising using a synergistic multi-resolution analysis of structural (MRI/CT) and functional datasets," *J. Nucl. Med.*, vol. 49, pp. 657–666, 2008.

[16] S. Tomei, S. Marache-Francisco, C. Odet, and C. Lartizien, "Automatic detection of active nodules in 3D PET oncology imaging using the hotelling observer and the support vector machines: A comparison study," in *IEEE NSS-MIC Conf. Rec.*, 2008, pp. 5314–5319.

[17] M. Hatt, C. Cheze Le Rest, A. Turzo, C. Roux, and D. Visvikis, "A fuzzy locally adaptive Bayesian segmentation approach for volume determination in PET," *IEEE Trans. Med. Imag.*, to be published.

[18] A. Reilhac, G. Batan, C. Michel, C. Grova, J. Tohka, D. L. Collins, N. Costes, and A. C. Evans, "PET-SORTEO: Validation and development of a database of simulated PET volumes," *IEEE Trans. Nucl. Sci.*, vol. 52, pp. 1321–1328, 2005.

[19] I. Castiglioni, I. Buvat, G. Rizzo, M. C. Girardi, J. Feuardent, and F. Fazio, "A publicly accessible Monte Carlo database for validation purpose in emission tomography," *Eur. J. Nucl. Med. Mol. Imag.*, vol. 32, pp. 1234–1239, 2005.

[20] I. G. Zubal, C. R. Harell, E. O. Smith, Z. Rattner, G. Gindi, and B. Hoffer, "Computerized three dimensional segmented human anatomy," *Phys. Med. Biol.*, vol. 21, no. 2, pp. 299–302, 1994.

[21] M. Aristophanous, B. C. Penney, and C. A. Pelizzari, "The development and testing of a digital PET phantom for the evaluation of tumor volume segmentation techniques," *Med. Phys.*, vol. 35, no. 7, pp. 3331–3342, 2008.

[22] S. Tomei, A. Reilhac, D. Visvikis, C. Odet, F. Giammarile, T. Mognetti, and C. Lartizien, "Development of a database of realistic simulated whole body 18F-FDG images for lymphoma," in *IEEE NSS-MIC Conf. Rec.*, 2008, pp. 4958–4963.

[23] J. A. Browne and A. R. Pierre, "A row-action alternative to the EM algorithm for maximizing likelihoods in emission tomography," *IEEE Trans. Med. Imag.*, vol. 15, pp. 687–699, 1996.

[24] D. Visvikis, A. Turzo, S. Gouret, P. Damine, F. Lamare, Y. Bizais, and C. Cheze Le Rest, "Characterisation of SUV accuracy in FDG PET using 3D RAMLA and the Philips Allegro PET scanner," *J. Nucl. Med.*, vol. 45, p. 103P, 2004.

[25] W. P. Segars, "Development of a new dynamic NURBS-based cardiac-torso (NCAT) phantom," Ph.D. dissertation, Univ. of North Carolina, Chapel Hill, May 2001.

[26] W. P. Segars, D. S. Lalush, and B. M. W. Tsui, "Development of an interactive software application to model patient populations in the spline-based MCAT phantom," in *IEEE NSS-MIC Conf. Rec.*, 2000, vol. 2, pp. 2051–2055.

[27] D. Ramos, Y. E. Erdi, M. Gonen, E. Riedel, H. W. D. Yeung, H. A. Macapinlac, R. Chisin, and S. M. Larson, "FDG-PET standardized uptake values in normal anatomical structures using iterative reconstruction segmented attenuation correction and filtered back-projection," *Eur. J. Nucl. Med. Mol. Imag.*, vol. 28, pp. 155–164, 2001.

[28] W. P. Segars, D. S. Lalush, and B. M. W. Tsui, "Modeling respiratory mechanics in the MCAT and spline-based MCAT phantoms," *IEEE Trans. Nucl. Sci.*, vol. 48, no. 1, pp. 89–97, 2001.

[29] S. A. Nehmeh, Y. E. Erdi, T. Pan, E. Yorke, G. S. Mageras, K. E. Rosenzweig, H. Schoder, H. Mostafavi, O. Squire, A. Pevsner, S. M. Larson, and J. Humm, "Quantitation of respiratory motion during 4D-PET/CT acquisition," *Med. Phys.*, vol. 31, pp. 1333–1338, 2004.

[30] M. E. Casey, H. Gadagkar, and D. Newport, "A component based method for normalization in volume PET," in *Proc. Int. Meeting Fully 3-D Image Reconstruct. Radiol. Nucl. Med.*, Aix-les-Bains, France, 1995, pp. 67–71.

[31] C. Watson, D. Newport, and M. E. Casey, "A single scatter simulation technique for scatter correction in 3D PET," in *Proc. Int. Meeting Fully 3-D Image Reconstruct. Radiol. Nucl. Med.*, P. Grangeat and J. L. Amans, Eds., 1996, pp. 225–268.

[32] W. P. Segars, M. Mahadevappa, T. Beck, E. C. Frey, and B. M. W. Tsui, "Validation of the 4D NCAT simulation tools for use in high resolution x-ray CT research," *J. Proc. SPIE*, vol. 5745, pp. 828–834, 2005.

[33] C. Burger, G. Goerres, S. Schoenes, A. Buck, A. H. Lonn, and G. K. Von Schulthess, "PET attenuation coefficients from CT images: Experimental evaluation of the transformation of CT into PET 511-keV attenuation coefficients," *Eur. J. Nucl. Med. Mol. Imag.*, pp. 922–927, 2002.

[34] C. Comtat, P. E. Kinahan, M. Defrise, C. Michel, C. Lartizien, and D. W. Townsend, "Simulating whole-body PET scanning with rapid analytical methods," in *IEEE NSS-MIC Conf. Rec.*, 1999, vol. 3, pp. 1260–1264.

[35] C. Michel, M. C. Sibomana, A. Bol, X. Bernard, M. Lonneux, M. Defrise, C. Comtat, P. E. Kinahan, and D. W. Townsend, "Preserving poisson characteristics of PET date with weighted OSEM reconstruction," in *IEEE NSS-MIC Conf. Rec.*, 1998, vol. 2, pp. 1323–1329.

[36] M. Soret, S. L. Bacharach, and I. Buvat, "Partial-volume effect in PET tumour imaging," *J. Nucl. Med.*, vol. 48, pp. 932–945, 2007.

[37] N. Boussion, M. Hatt, F. Lamare, Y. Bizais, A. Turzo, C. Cheze-Le Rest, and D. Visvikis, "A multiresolution image based approach for correction of partial volume effects in emission tomography," *Phys. Med. Biol.*, vol. 51, no. 7, pp. 1857–1876, 2006.

[38] J. Tohka and A. Reilhac, "Deconvolution-based partial volume correction in raclopride-PET and Monte Carlo comparison to MR-based method," *Neuroimage*, vol. 39, pp. 1570–1584, 2008.

[39] B. K. Teo, Y. Seo, S. L. Bacharach, J. A. Carrasquillo, S. K. Libutti, H. Shukla, B. H. Hasegawal, R. A. Hawkins, and B. L. Franc, "Partial-volume correction in PET: Validation of an iterative postreconstruction method with phantom and patient data," *J. Nucl. Med.*, vol. 48, no. 5, pp. 802–810, 2007.

[40] A. S. Kirov, J. Z. Piao, and C. R. Schmidtlein, "Partial volume effect correction in PET using regularized iterative deconvolution with variance control based on local topology," *Phys. Med. Biol.*, vol. 53, pp. 2577–2591, 2008.

[41] O. G. Rousset, Y. Ma, and A. C. Evans, "Correction for partial volume effects in PET: Principle and validation," *J. Nucl. Med.*, vol. 39, no. 5, pp. 904–911, 1998.

[42] A. Sovik, E. Malinen, and D. G. Olsen, "Strategies for biologic image-guided dose escalation: A review," *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 73, no. 3, pp. 650–658, 2009.

[43] F. O'Sullivan, S. Roy, and J. Eary, "A statistical measure of tissue heterogeneity with application to 3D PET sarcoma data," *Biostatistics*, vol. 4, pp. 433–448, 2003.

[44] W. P. Segars, B. M. W. Tsui, E. C. Frey, and E. K. Fishman, "Extension of the 4D NCAT phantom to dynamic X-ray CT simulation," in *IEEE NSS-MIC Conf. Rec.*, 2004, vol. 5, pp. 3195–3199.

[45] P. Descourt, W. P. Segars, F. Lamare, L. Ferrer, B. M. W. Tsui, Y. Bizais, M. Bardies, and D. Visvikis, "RTNCAT (real time NCAT): Implementing real time physiological movement of voxellised phantoms in GATE," in *IEEE NSS-MIC Conf. Rec.*, 2006, vol. 5, pp. 3163–3165.

[46] F. Lamare, M. J. L. Carbayo, T. Cresson, G. Kontaxakis, A. Santos, C. Cheze Le Rest, A. J. Reader, and D. Visvikis, "List mode based image reconstruction for respiratory motion correction in PET using non-rigid body transformations," *Phys. Med. Biol.*, vol. 52, pp. 5187–5204, 2007.

## ABOUT THE AUTHORS

**Amandine Le Maitre** received the engineering and master's degrees in medical image processing from Telecom Bretagne, France, in 2008. She is currently pursuing the Ph.D. degree in the Medical Image Processing Lab (LaTIM, INSERM U650), Brest, France.

Her current research interests include the incorporation of patient specific variability in the simulation of emission tomography datasets for oncology applications and modeling of tumor processes for outcome prediction in radiotherapy treatment planning.

**William Paul Segars** (Member, IEEE) received the Ph.D. degree in biomedical engineering from the University of North Carolina, Chapel Hill, in 2001.

He is an Assistant Professor of radiology and biomedical engineering and a Member of the Carl E. Ravin Advanced Imaging Laboratories (RAILabs) at Duke University, Durham, NC. He is among the leaders in the development of simulation tools for medical imaging research, where he has applied state-of-the-art computer graphics techniques to develop realistic anatomical and physiological models. Foremost among these are the extended 4-D NURBS-based cardiac-torso (XCAT) phantom, a computational model for the human body, and the 4-D mouse whole-body (MOBY) phantom, a model for the laboratory mouse. These phantoms are widely used to evaluate and improve imaging devices and techniques.

**Simon Marache** received the bachelor of information technology degree and the master's degree in pattern recognition from the National Institute for Applied Sciences, Rouen, France, in 2008. He is currently pursuing the Ph.D. degree at the University of Lyon, France.

He is with the CREATIS laboratory, whose aim is to develop image-processing methods for medical imaging. His thesis centers around performance estimation of motion correction techniques for tumor detection in 3-D PET/CT medical imaging.

**Anthonin Reilhac**, photograph and biography not available at the time of publication.

**Mathieu Hatt** (Member, IEEE) received the master's degree in computer sciences from the Université de Strasbourg, France, and the Ph.D. degree in image processing from the Université de Bretagne Occidentale, Brest, France, in 2008.

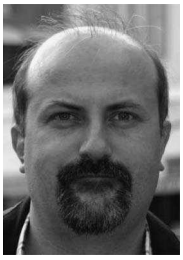
He is currently an R&D Engineer with the Medical Image Processing Lab (LaTIM, INSERM U650), Brest. His research interests include unsupervised image segmentation and classification, fuzzy logic, Bayesian and Markovian modeling, and multimodality imaging for oncology applications.

**Sandrine Tomei** received the bachelor's degree in computing and electronic sciences from Chimie Physique Électronique de Lyon, France, and the master's degree in image and signal processing from INSA-Lyon, France. She is currently pursuing the Ph.D. degree at the University of Lyon.

She is conducting her doctoral research at the CREATIS-LRMN lab, focusing on automatic hot-spot detection in whole-body PET images.

**Carole Lartizien** received the bachelor's degree in nuclear engineering from the National Polytechnic Institute, Grenoble, France, in 1996. She received the master's degree in biomedical engineering and the Ph.D. degree in medical imaging from the University de Paris XI, France, in 1997 and 2001, respectively.

She is a Research Associate with CNRS and the University of Lyon, France, and is conducting research at the CREATIS laboratory, whose aim is to develop image-processing methods for medical imaging. Her current research centers around methodology for estimation and detection tasks in 3-D PET imaging and simulation of medical images.



**Dimitris Visvikis** (Senior Member, IEEE) is a Senior Research Scientist with the National Institute of Health and Medical Sciences (INSERM), France. He is based within the Medical Image Processing Lab in Brest (LaTIM, INSERM U650), where he is in charge of a group on quantitation in multimodality imaging for diagnostic and therapy applications. He has spent the majority of his scientific activity in the field of PET imaging, including developments in both hardware and software domains. His current research interests focus on improvement in PET/CT image quantitation for specific oncology applications, such as response to therapy and radiotherapy treatment planning, through the development of methodologies for detection and correction of respiratory motion, partial volume correction, and tumor volume segmentation algorithms, as well as the development and validation of Monte Carlo simulations for emission tomography applications.

Dr. Visvikis is a member of the SNM Computer and Instrumentation Council and a past member of the IEEE NMISC board.

# Intratumor Heterogeneity Characterized by Textural Features on Baseline $^{18}$F-FDG PET Images Predicts Response to Concomitant Radiochemotherapy in Esophageal Cancer

Florent Tixier[1], Catherine Cheze Le Rest[1,2], Mathieu Hatt[1], Nidal Albarghach[1,3], Olivier Pradier[1,3], Jean-Philippe Metges[3,4], Laurent Corcos[4], and Dimitris Visvikis[1]

*[1]INSERM, U650, LaTIM, CHU Morvan, Brest, France; [2]Department of Nuclear Medicine, CHU Morvan, Brest, France; [3]Institute of Oncology, CHU Morvan, Brest, France; and [4]INSERM, U613, Faculty of Medicine, Brest, France*

$^{18}$F-FDG PET is often used in clinical routine for diagnosis, staging, and response to therapy assessment or prediction. The standardized uptake value (SUV) in the primary or regional area is the most common quantitative measurement derived from PET images used for those purposes. The aim of this study was to propose and evaluate new parameters obtained by textural analysis of baseline PET scans for the prediction of therapy response in esophageal cancer. **Methods:** Forty-one patients with newly diagnosed esophageal cancer treated with combined radiochemotherapy were included in this study. All patients underwent pretreatment whole-body $^{18}$F-FDG PET. Patients were treated with radiotherapy and alkylatinlike agents (5-fluorouracil-cisplatin or 5-fluorouracil-carboplatin). Patients were classified as nonresponders (progressive or stable disease), partial responders, or complete responders according to the Response Evaluation Criteria in Solid Tumors. Different image-derived indices obtained from the pretreatment PET tumor images were considered. These included usual indices such as maximum SUV, peak SUV, and mean SUV and a total of 38 features (such as entropy, size, and magnitude of local and global heterogeneous and homogeneous tumor regions) extracted from the 5 different textures considered. The capacity of each parameter to classify patients with respect to response to therapy was assessed using the Kruskal–Wallis test ($P < 0.05$). Specificity and sensitivity (including 95% confidence intervals) for each of the studied parameters were derived using receiver-operating-characteristic curves. **Results:** Relationships between pairs of voxels, characterizing local tumor metabolic nonuniformities, were able to significantly differentiate all 3 patient groups ($P < 0.0006$). Regional measures of tumor characteristics, such as size of nonuniform metabolic regions and corresponding intensity nonuniformities within these regions, were also significant factors for prediction of response to therapy ($P = 0.0002$). Receiver-operating-characteristic curve analysis showed that tumor textural analysis can provide nonresponder, partial-responder, and complete-responder patient identification with higher sensitivity (76%–92%) than any SUV measurement. **Conclusion:** Textural features of tumor metabolic distribution extracted from baseline $^{18}$F-FDG PET images allow for the best stratification of esophageal carcinoma patients in the context of therapy-response prediction.

**Key Words:** $^{18}$F-FDG PET; esophageal cancer; textural analysis; predictive value; response to therapy

**J Nucl Med 2011; 52:369–378**
DOI: 10.2967/jnumed.110.082404

Esophageal cancer is associated with high mortality. In patients with early-stage disease at presentation, esophagectomy is the treatment of choice and is potentially curative. Unfortunately most patients at presentation have already locally advanced esophageal cancer or distant metastases. In locally advanced esophageal cancer, preoperative chemotherapy or radiochemotherapy will improve survival in patients who respond to induction therapy (*1,2*). On the other hand, patients who do not respond to neoadjuvant therapy may be affected unnecessarily by the toxicity of an inefficient therapy. Therefore, the development of a diagnostic test capable of noninvasively predicting response to therapy early in the course of treatment is of great interest, potentially allowing personalization of patient management. In patients treated by exclusive conventional combined radiochemotherapy, assessment of response is equally of great interest, because it could allow an early change in the management of nonresponding patients. Such assessment becomes more critical when one considers the availability of new targeted therapies that could be tested with higher efficiency if applied early in diagnosis (*3,4*).

$^{18}$F-FDG PET is already well established for the initial staging of esophageal cancer, because it is associated with a better sensitivity and specificity than combined use of CT and echoendoscopy, especially regarding detection of distant metastasis (*5*).

$^{18}$F-FDG PET has been also used to assess response to therapy and patient outcome prognosis (*4,6*). Within this context, few studies have explored the potential prognostic value of pretreatment $^{18}$F-FDG PET, demonstrating that the level of activity concentration on preoperative PET,

Received Aug. 17, 2010; revision accepted Dec. 13, 2010.
For correspondence or reprints contact: Florent Tixier, LaTIM, INSERM U650, CHU Morvan, 5 Avenue Foch, 29609 Brest, France.
E-mail: florent.tixier@etudiant.univ-brest.fr

although not statistically significant, tends to predict overall survival (*7–9*).

On the other hand, several studies have evaluated the role of PET in assessing treatment response based on $^{18}$F-FDG uptake changes between a pre- and a posttreatment PET scan obtained during or after the treatment completion. Studies considering a second PET scan after treatment completion have shown that a complete metabolic response is associated with better outcome (*10–12*). However, because that information is of limited interest in patient management if acquired late, different attempts have been made to determine whether $^{18}$F-FDG PET could be used for assessing response to therapy earlier (usually within a few weeks) in the course of treatment (*13–15*), showing some promising results that need to be confirmed in multicenter trials (*4*). One of the highlighted issues is that early response prediction during combined chemoradiotherapy, in contrast to chemotherapy alone, may be compromised by increased $^{18}$F-FDG tumor uptake resulting from radiotherapy-induced inflammatory processes (*4*).

An alternative to monitoring changes during treatment is the potential of predicting response to therapy from the baseline $^{18}$F-FDG PET scan alone, which may allow the best available therapy regime to be chosen for a given patient. However, to date there is only limited evidence that a measure of tumor activity concentration on a baseline PET scan in esophageal cancer can differentiate groups of patient response (*8,9*). Within the same context, parameters derived from pretreatment $^{18}$F-FDG PET have shown the potential to differentiate between responders and nonresponders (NRs) in non-Hodgkin lymphoma patients (*16*).

The PET image index predominantly used in such studies for assessment of metabolic response is the normalized mean tumor activity concentration known as the mean standardized uptake value ($SUV_{mean}$), within a region of interest around the tumor, or the maximum standardized uptake value corresponding to the highest-activity pixel value ($SUV_{max}$). However, $^{18}$F-FDG tumor uptake has been associated not only with increased metabolism but also with several other physiologic parameters such as perfusion, cell proliferation (*17*), tumor viability, aggressiveness, or hypoxia (*18,19*), all of which may in turn be responsible for tumor uptake heterogeneity. Therefore, the hypothesis can be made that characterizing tumor $^{18}$F-FDG distribution, through its relationship to underlying tumor biologic characteristics, may be useful in predicting therapy response. $^{18}$F-FDG tumor activity distribution may be assessed in a global, regional, or local fashion, allowing in turn the assessment of corresponding global, regional, or local patterns of biologic heterogeneity. Although the measurement of such features have been previously explored in anatomic imaging (*20–22*), they have not to date been widely used in PET. Until now, only 1 study has considered the use of some textural features to predict treatment outcome from baseline $^{18}$F-FDG PET images, with encouraging results in cervical and head and neck cancer (*23*), and the assessment of spatial heterogeneity was also shown to be significantly associated with survival in sarcoma patients (*24*). However, the potential predictive value of tumor heterogeneity characterization on a baseline $^{18}$F-FDG PET scan has never been assessed.

The objective of this current study was, therefore, to assess the predictive value of $^{18}$F-FDG uptake heterogeneity characterized by textural features extracted from pretherapy $^{18}$F-FDG PET images of patients with esophageal carcinoma by assessing the ability of each parameter to identify different categories of responders. The predictive value of these parameters was compared with the use of standard image activity concentration indices ($SUV_{max}$, $SUV_{mean}$). The potential prognostic value of such image-derived parameters for assessing overall patient survival was not assessed in this study.

## MATERIALS AND METHODS

### Patients

Forty-one patients with newly diagnosed esophageal cancer treated with exclusive radiochemotherapy between 2003 and 2008 were included in this study. The characteristics of the patients are summarized in Table 1. The mean age at the time of diagnosis was 66 ± 10 y (median, 69 y; range, 45–84 y), and 85% of patients were male. Most of the tumors were squamous cell carcinoma (76%), and most of the patients had a well or moderately differentiated tumor (56%). Most of the tumors originated from the middle and lower esophagus (76%). Twenty-six patients had a T3 or T4 primary lesion, 25 had N1 (61%) lymph node metastases, and 17 had distant metastases (Table 1). All patients were treated with external-beam radiotherapy and chemotherapy with alkylatin-like agents (5-fluorouracil-cisplatin or 5-fluorouracil-carboplatin). A median radiation dose of 60 Gy was delivered in 180-cGy daily fractions (5 d/wk and 6–7 wk in total). One month after the completion of the treatment, patients were reassessed to determine response to therapy using thoracoabdominal CT and endoscopy. Patients were subsequently classified as complete responders (CR), partial responders (PR), stable disease, or progressive disease. Response was assessed using pretreatment and posttreatment CT scans by evaluating the increase (or decrease) in the sum of the longest diameters for all target lesions and the appearance, persistence, or disappearance of nontarget lesions, according to the Response Evaluation Criteria in Solid Tumors (RECIST) (*25*). Considering the small number of patients in the stable disease (7) and progressive disease (4) groups, these patients were eventually combined into an NR group.

All patients underwent pretreatment whole-body $^{18}$F-FDG PET for staging purposes. Patients were instructed to fast for a minimum of 6 h before the injection of $^{18}$F-FDG. The dose of administered $^{18}$F-FDG was 5 MBq/kg, and static emission images were acquired from thigh to head, on average 54 min after injection, on a Gemini PET/CT scanner (Philips). In addition to the emission PET scan, a low-dose CT scan was acquired for attenuation-correction purposes. Images were reconstructed with the 3-dimensional (3D) row-action maximum-likelihood algorithm using standard clinical protocol parameters (2 iterations, relaxation parameter of 0.05, and 3D gaussian postfiltering of 5 mm in

**TABLE 1**
Characteristics of Patients ($n$ = 41)

| Characteristic | No. of patients |
|---|---|
| Sex | |
| Male | 35 (85) |
| Female | 6 (15) |
| Primary site | |
| Upper esophagus | 10 (24) |
| Middle esophagus | 15 (37) |
| Lower esophagus | 16 (39) |
| Tumor cell type | |
| Squamous cell carcinoma | 31 (76) |
| Adenocarcinoma | 10 (24) |
| Histologic grade | |
| Well differentiated | 12 (29) |
| Moderately differentiated | 11 (27) |
| Poorly differentiated | 3 (7) |
| Unknown | 15 (37) |
| TNM stage | |
| T1 | 6 (15) |
| T2 | 7 (17) |
| T3 | 21 (51) |
| T4 | 7 (17) |
| N0 | 16 (39) |
| N1 | 25 (61) |
| M0 | 24 (59) |
| M1 | 17 (41) |
| AJCC stage | |
| I | 4 (10) |
| IIa | 6 (15) |
| IIb | 5 (12) |
| III | 12 (29) |
| IVa | 4 (10) |
| IVb | 10 (24) |
| RECIST | |
| CR | 9 (22) |
| PR | 21 (51) |
| Stable disease (NR) | 7 (17) |
| Progressive disease (NR) | 4 (10) |

Data in parentheses are percentages.

full width at half maximum). The current data analysis was performed after approval by the institutional review board.

### Tumor Analysis

For each patient, primary tumors were identified on $^{18}$F-FDG PET images by an experienced nuclear physician. Tumors were then delineated automatically using the previously validated fuzzy locally adaptive Bayesian algorithm (*26*). All parameters were subsequently extracted from this delineated volume. Only the primary tumors were considered because textural analysis cannot be reliably performed on small lesions (nodal or distant metastases) because of the small number of voxels involved.

### Standardized Uptake Value (SUV) Analysis

The following SUV parameters were extracted from each patient's baseline PET images: $SUV_{max}$; peak SUV ($SUV_{peak}$), defined as the mean of the voxel of maximum value and its 26 neighbors (in 3 dimensions); and mean SUV within the delineated tumor ($SUV_{mean}$). The $SUV_{peak}$ was considered in addition to $SUV_{max}$ to investigate the impact of reducing the potential bias in the $SUV_{max}$ measurements as a result of its sensitivity to noise.

### Textural Analysis

We define texture as a spatial arrangement of a predefined number of voxels allowing the extraction of complex image properties, and we define a textural feature as a measurement computed using a texture matrix. The method used was realized in 2 steps. First, matrices describing textures on images were extracted from tumors, and textural features were subsequently computed using theses matrices. All these parameters characterize in some way tumor heterogeneity at local and regional (using texture matrices) or global scales (using image-voxel-intensity histograms).

Several different textures (Table 2, left column) were computed. Voxel values within the segmented tumors (Fig. 1A and 1B) were resampled to yield a finite range of values (Fig. 1C), allowing textural analysis using:

$$V(x) = \left[ 2^s \frac{I(x) - \min\limits_{i\in\Omega} i}{\max\limits_{i\in\Omega} i - \min\limits_{i\in\Omega} i + 1} \right] \quad \text{Eq. 1}$$

where $2^S$ represents the number of discrete values (16–128), $I$ is the intensity of the original image, and $\Omega$ is the set of voxels in the delineated volume. This resampling step on the delineated tumor volume, necessary for the computation of the textural analysis, has 2 effects: it reduces the noise in the image by clustering voxels with similar intensities and it normalizes the tumor voxel intensities across patients, which in turn facilitates the comparison of the textural features. Local and regional features were computed with different resampling considering 16, 32, 64, and 128 discrete values to investigate the potential impact of this resampling parameter.

All considered textures were originally described for 2 dimensions (*27–30*) and were therefore adapted in this work for 3 dimensions. The cooccurrence matrix (M1, Fig. 1D(a)) describing pairwise arrangement of voxels, and the matrix describing the alignment of voxels with the same intensity (M2, Fig. 1D(b)), were computed considering 13 different angular directions. Finally, 3D matrices describing differences between each voxel and its neighbors (M3, Fig. 1D(c)) and characteristics of homogeneous zones (M4, Fig. 1D(d)) were computed considering for each voxel the neighbors in the 2 adjacent planes, adapting the normalizing factors to 3 dimensions.

From each of the extracted texture matrices, different features summarized in Table 2 (middle column) were computed. Depending on the way the matrix is analyzed, it is possible to extract features of a local or regional nature. Six features highlighting local variations of voxel intensities within the image were extracted from the cooccurrence matrices M1 (Fig. 2C). For example, using the matrix M1, the local entropy and homogeneity are calculated using Equations 2 and 3, respectively:

$$\text{Local entropy} = -\sum_{i,j} M1(i,j) log(M(i,j)) \quad \text{Eq. 2}$$

$$\text{Local homogeneity} = \sum_{i,j} \frac{M1(i,j)}{1+|i-j|} \quad \text{Eq. 3}$$

where M1 is a cooccurrence matrix, *i, j* are the rows and columns index, and M1(*i,j*) is an element of the matrix.

In addition, M3 matrices were used to extract busyness (quantifying sharp-intensity variations) and contrast and coarse-

**TABLE 2**
Texture Type and Associated Features

| Type | Feature | Scale |
|---|---|---|
| Features based on intensity histogram | Minimum intensity<br>Maximum intensity<br>Mean intensity<br>Variance<br>SD<br>Skewness<br>Kurtosis | Global |
| Features based on voxel-alignment matrix (M2) | Short run emphasis<br>Long run emphasis<br>Intensity variability<br>Run-length variability<br>Run percentage<br>Low-intensity run emphasis<br>High-intensity run emphasis<br>Low-intensity short-run emphasis<br>High-intensity short-run emphasis<br>Low-intensity long-run emphasis<br>High-intensity long-run emphasis | Regional |
| Features based on intensity–size–zone matrix (M4) | Short-zone emphasis<br>Large-zone emphasis<br>Intensity variability<br>Size-zone variability<br>Zone percentage<br>Low-intensity zone emphasis<br>High-intensity zone emphasis<br>Low-intensity short-zone emphasis<br>High-intensity short-zone emphasis<br>Low-intensity large-zone emphasis<br>High-intensity large-zone emphasis | Regional |
| Features based on cooccurrence matrices (M1) | Second angular moment<br>Contrast (inertia)<br>Entropy<br>Correlation<br>Homogeneity<br>Dissimilarity | Local |
| Features based on neighborhood intensity-difference matrix (M3) | Coarseness<br>Contrast<br>Busyness | Local |

ness (quantifying tumor granularity). These features allow extracting measurements describing tumor local heterogeneity proportional to variations of $^{18}$F-FDG uptake between individual voxels.

On the other hand, the M2 and M4 matrices were used to extract regional tumor uptake characteristics, representing regional heterogeneity, such as variation of intensity between regions and in the size and alignment of homogeneous areas. For example, the M4 matrix links the homogeneous tumor regions to their intensity (Fig. 2B). It was hence used to calculate the variability in the size and the intensity of identified homogeneous tumor zones according to Equations 4 and 5, respectively:

$$\text{Size-zone variability} = \frac{1}{\Theta}\sum_{m=1}^{M}\left[\sum_{n=1}^{N} M4(m,n)\right]^2 \qquad \text{Eq. 4}$$

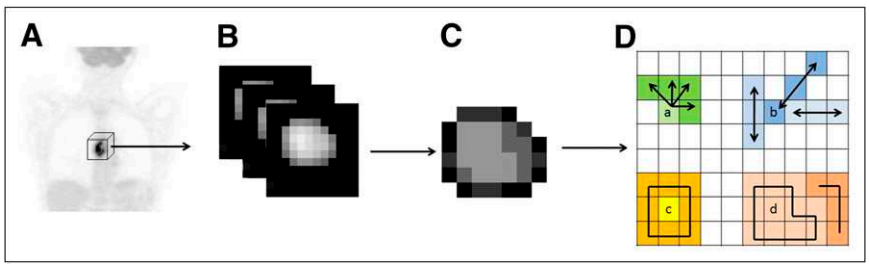



**FIGURE 1.** Whole-body $^{18}$F-FDG PET scan (A), tumor segmentation (B), and voxel-intensity resampling (C) allowing extraction of different features (D) by analysis of consecutive voxels in a direction (for cooccurrence matrices) (a), alignment of voxels with same intensity (b), difference between voxels and their neighbors (c), and zones of voxels with same intensity (d).

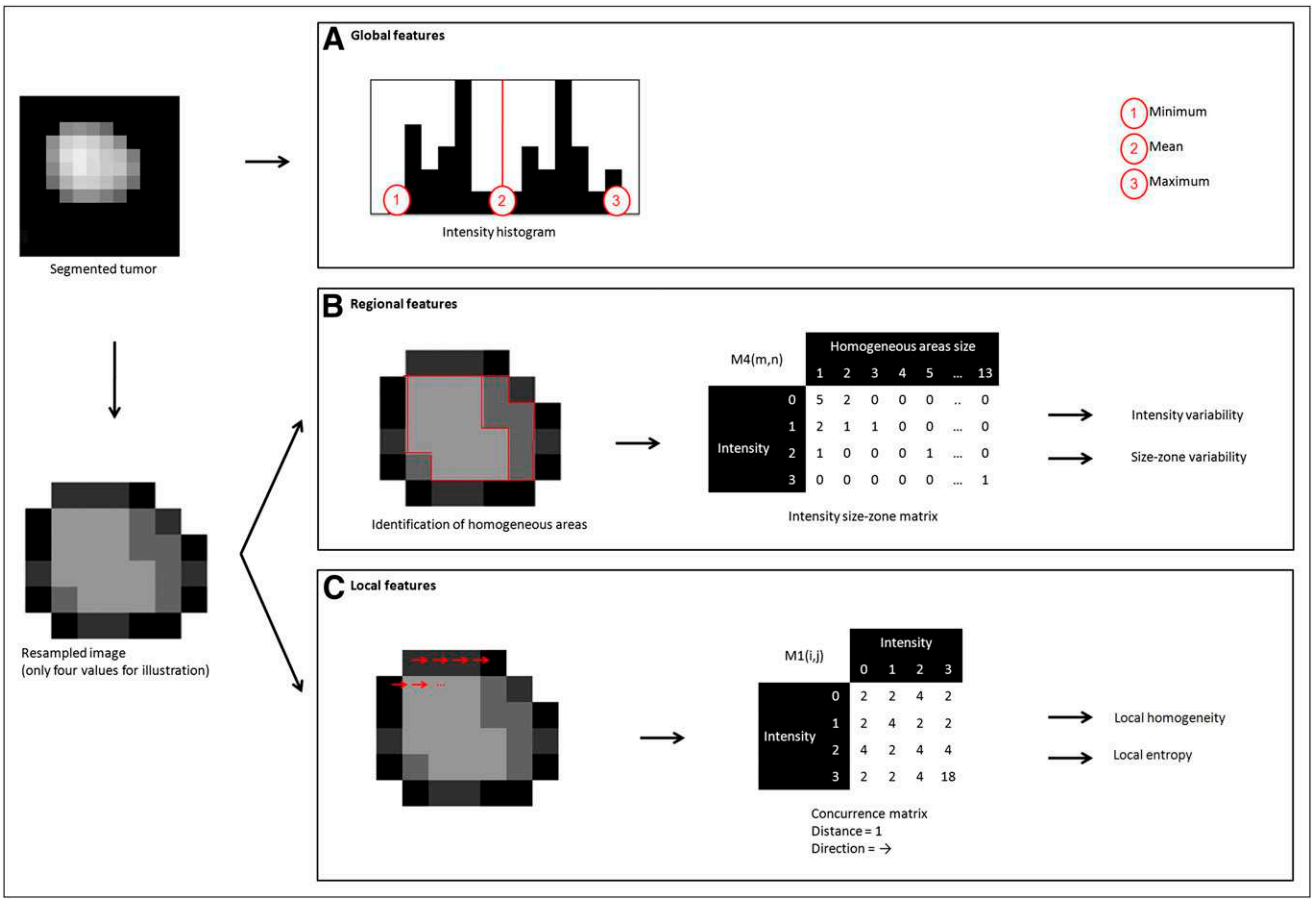



**FIGURE 2.** Examples of features extracted from tumor resampled on 4 values: 3 global features computed using intensity histogram, 2 regional features computed using M4 matrix, and 2 local features computed using M1 texture matrices.

$$\text{Intensity variability} = \frac{1}{\Theta}\sum_{n=1}^{N}\left[\sum_{m=1}^{M} M4(m,n)\right]^2, \qquad \text{Eq. 5}$$

where $\Theta$ represents the number of homogeneous areas in the resampled tumor, $M$ the number of distinct intensity values within the tumor, and $N$ the size of the largest homogeneous area in the matrix M4.

Finally, global features are computed on the original image voxels' intensity distribution by analyzing the characteristics of the intensity value histogram within the segmented tumor (Fig. 2A).

Thirty-eight features were extracted from the 4 different texture matrices and intensity histograms. Seven of the 38 features characterize the uptake distribution within the entire tumor (using the intensity histogram), 9 describe local voxel arrangements (using matrices M1 and M3), and 22 are related to the organization of voxels at a regional scale (using matrices M2 and M4).

### Statistical Analysis

The capacity of each feature to classify patients with respect to therapy response was investigated on the primary tumor using the Kruskal–Wallis test (*8*). $P$ values of less than 0.05 were considered statistically significant. Specificity and sensitivity (including 95% confidence intervals [CIs]) for each of the studied parameters were derived using receiver operating characteristic (ROC) curves measuring associated areas under the ROC curves (AUC). Texture results were compared with those of $SUV_{max}$, $SUV_{mean}$, and $SUV_{peak}$ for their ability to distinguish among responders (PR and CR) and NRs, CRs and non-CRs (PR, NR), and all 3 groups separately.

## RESULTS

Patients were evaluated 1 mo after the completion of combined radiochemotherapy. Nine patients (22%) had no evidence of disease after treatment and were considered CRs. Radiochemotherapy led to partial response in 21 (51%) patients, whereas 11 (27%) were stable or progressed under treatment according to RECIST (*25*).

Results of the Kruskal–Wallis test show that $SUV_{max}$ (Fig. 3) and $SUV_{mean}$ were capable of differentiating only CRs from NRs and PRs. Within this context, all SUV measurements were significant predictive factors of response ($P = 0.034$, 0.044, and 0.012 for the $SUV_{max}$, $SUV_{mean}$, and $SUV_{peak}$, respectively). However, only $SUV_{peak}$ was a significant predictive factor ($P = 0.045$) when considering the differentiation of 3 patient response groups (i.e., NR, PR, and CR), whereas $SUV_{max}$ and $SUV_{mean}$ were not ($P > 0.05$).

Figure 4 shows examples of different extracted features and associated values for tumors of CRs, PRs, and NRs. The Kruskal–Wallis tests revealed no statistically significant differences in the textural parameters derived using

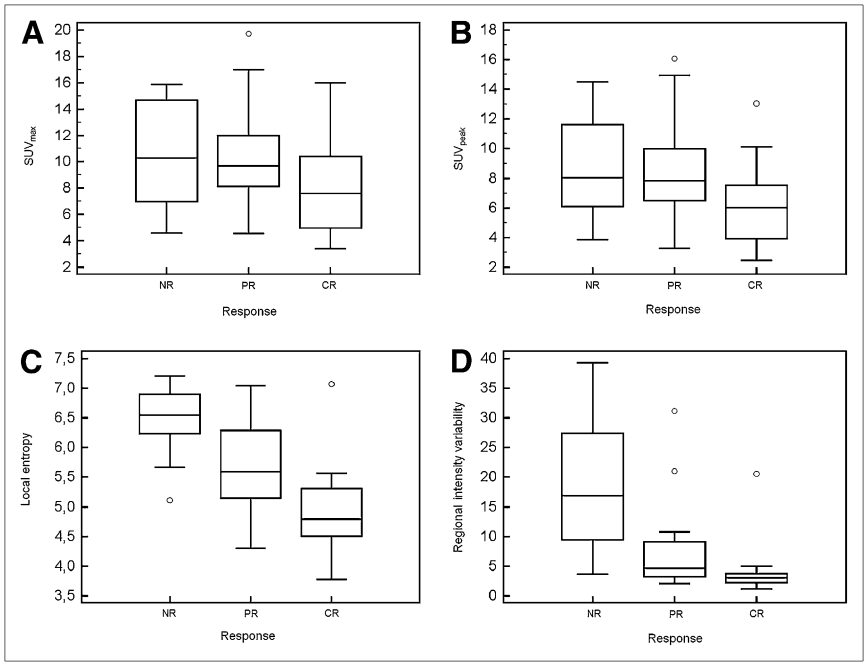



**FIGURE 3.** Box-plot representation of parameters' values in function of patient response (0, NR; 1, PR; and 2, CR) for $SUV_{max}$ ($P = 0.106$) (A), $SUV_{peak}$ ($P = 0.045$) (B), local entropy ($P = 0.0006$) (C), and regional intensity variability ($P = 0.0002$) (D).

different resampling values (16, 32, 64, or 128 discrete values). All subsequent reported results were obtained using 64 discrete values in the resampling normalization process. This value was chosen because it allows for 0.25 SUV increments, which were considered sufficient given the range of SUVs encountered (from ~4–20).

None of the global features extracted from the intensity histogram within the tumor was a significant predictive factor of response to therapy. However, considering local variation of $^{18}$F-FDG uptake, a high predictive value ($P < 0.0007$) was found using the cooccurrence features, particularly considering the use of the average feature values

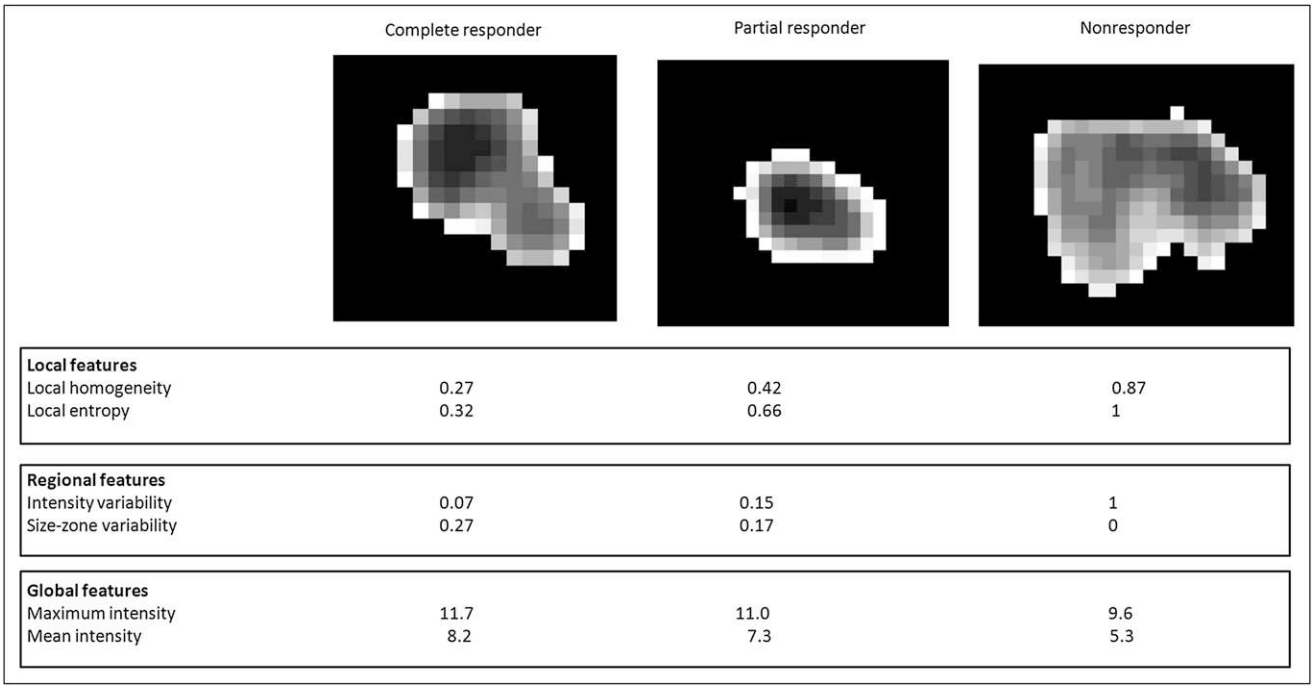



**FIGURE 4.** Example of different extracted features and associated values for tumors of CRs, PRs, and NRs (results are normalized in [0–1] interval using range of observed values for local and regional parameters).

computed using M1 matrices (Table 3). All these features offered statistically significant differentiation of NRs and responders (considering both CR and PR).

Regarding local features, the busyness and contrast computed on M3 matrices were not statistically significant predictive factors of response, but the coarseness, reflecting the local granularity of the tumor functional image, was found to be significant ($P = 0.0002$). Among the local measures of functional tumor characteristics computed using M1 matrices, the measure of local entropy was the only measure allowing statistically significant differentiation of all 3 patient groups ($P = 0.0006$, Fig. 3).

Because the features computed on M2 and M4 matrices, used to highlight regional variability in the $^{18}$F-FDG distribution, were strongly correlated ($r > 0.9$), only features based on M4 were used in the subsequent analysis. Regional measures of tumor characteristics extracted from these M4 matrices, such as the variability in the size and the intensity of identified homogeneous tumor zones, were statistically significant in predicting therapy response ($P = 0.0002$), allowing the differentiation of all 3 patient response groups (Fig. 3).

The ROC curve analysis for $SUV_{max}$, $SUV_{peak}$, local homogeneity, local entropy, and regional tumor characteristics such as the variability in size and intensity of identified homogeneous tumor areas is presented in Figure 5. Table 3 summarizes the ROC curve analysis results, comparing the performance of the different studied parameters in terms of sensitivity and specificity in, on the one hand, identifying complete-response patients and, on the other hand, differentiating responders (PR and CR).

First, based on the ROC curve analysis, textural parameters can identify CRs better than can the SUV-based measurements, as demonstrated by the respective AUCs (Fig. 5). For example, $SUV_{max}$, with an AUC of 0.7, allowed the identification of CRs, with a maximum sensitivity of 46% and specificity of 91%, using a threshold of 6. On the other hand, the variability in the size of the uniform tumor zones (AUC, 0.85) allowed for the extraction of CR patients with the best accuracy (sensitivity, 92%; specificity, 69%).

Second, as Figure 5 shows, textural features were most efficient in identifying responders (CRs and PRs), whereas for the same task the performance of SUV measurements was limited. For the differentiation of the patient responders, the AUC was less than 0.6 for the different SUV parameters, compared with an AUC of more than 0.82 for the use of the texture parameters. For example, the AUC of the $SUV_{max}$ was 0.59, allowing a sensitivity of 53% and specificity of 73% in the differentiation of responders using an optimal threshold of 9.1. On the other hand, for the same task the local homogeneity had a specificity and sensitivity of 88% and 73%, respectively (AUC, 0.89).

## DISCUSSION

Assessment of tumor response to therapy plays a central role in drug development and patient clinical management. Currently, response is mainly assessed by measuring anatomic tumor size and classifying tumor shrinkage according to standard criteria. Because metabolic changes often occur before morphologic changes, metabolic imaging appears to be a valuable tool for monitoring various treatments in different cancer types. Within this context, $^{18}$F-FDG PET has shown promising results in assessing response to therapy and prognosis. In esophageal cancer, quantitative changes in $^{18}$F-FDG uptake at 2 wk after the start of therapy have been shown to correlate well with

**TABLE 3**
Sensitivity and Specificity (Along with Corresponding 95% Confidence Intervals) of 3 SUV-Based Measurements, 2 Cooccurrence Features, and 2 Size-Zone Features

| Comparison | Parameters | Sensitivity (%) | 95% confidence interval (%) | Specificity (%) | 95% confidence interval (%) |
|---|---|---|---|---|---|
| NR vs. PR + CR | $SUV_{max}$ | 53 | 35.1–70.2 | 73 | 39.0–94.0 |
| | $SUV_{mean}$ | 71 | 52.5–84.9 | 45 | 16.7–76.6 |
| | $SUV_{peak}$ | 56 | 37.9–72.8 | 73 | 39.0–94.0 |
| | Local homogeneity | 88 | 71.8–96.6 | 73 | 39.0–94.0 |
| | Local entropy | 79 | 61.1–91.0 | 82 | 48.2–97.7 |
| | Size-zone | 76 | 58.8–89.8 | 91 | 58.7–99.8 |
| | Intensity variability | 76 | 58.7–89.3 | 91 | 58.7–99.8 |
| NR + PR vs. CR | $SUV_{max}$ | 46 | 19.2–74.9 | 91 | 75.0–98.0 |
| | $SUV_{mean}$ | 62 | 31.6–86.1 | 81 | 63.6–92.8 |
| | $SUV_{peak}$ | 62 | 31.6–86.1 | 81 | 63.6–92.8 |
| | Local homogeneity | 92 | 61.5–99.8 | 56 | 37.7–73.6 |
| | Local entropy | 92 | 61.5–99.8 | 69 | 50.0–83.9 |
| | Size-zone | 92 | 64.0–99.8 | 69 | 50.0–83.9 |
| | Intensity variability | 85 | 54.6–98.1 | 75 | 56.6–88.5 |

Data in top part of table are evaluation of parameters to distinguish PR or CR; data on bottom part of table are evaluation of parameters to differentiate CRs.

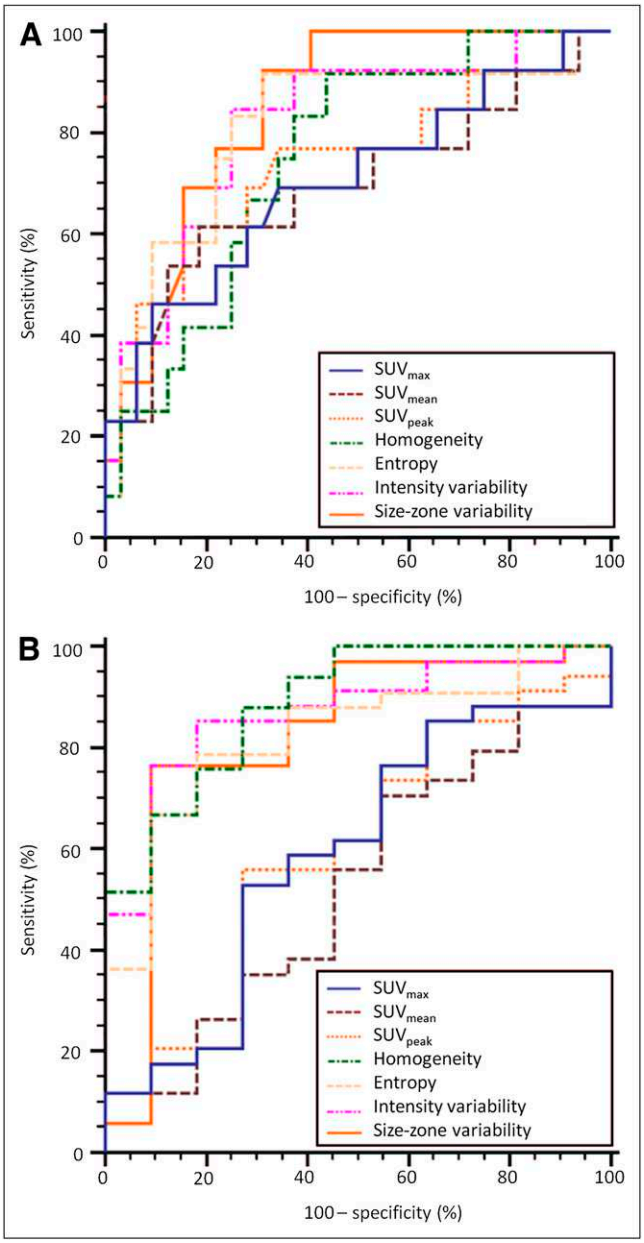



**FIGURE 5.** ROC curves for $SUV_{max}$, $SUV_{mean}$, $SUV_{peak}$, local homogeneity, uniform tumor areas, intensity variability, and size-zone variability for identification of CRs (A) and PRs or CRs (B).

subsequent tumor shrinkage and patient survival (*4*). This approach still has limitations, especially if patients undergo radiotherapy treatment. Hautzel et al. have shown that even low irradiation may enhance tumor uptake, and inflammatory changes may contribute early to this increase, yielding inaccurate information about treatment response (*31*). Within the same context, induced ulceration may also impair response assessment using PET (*32*).

On the other hand, the prediction of response before treatment initiation may be of great interest to the optimization of patient management. With such an endpoint, few authors have studied the predictive value of initial $^{18}$F-FDG uptake for therapy response. Rizk et al. reported an $SUV_{max}$ of more than 4.5 to be a reliable predictor of pathologic response (*9*), whereas Javeri et al. (*8*) demonstrated in a larger group of patients a trend of greater rate of response obtained after combined chemoradiotherapy in patients who had an initial $SUV_{max}$ higher than 10. Similarly in our study, initial $SUV_{mean}$, $SUV_{max}$, and $SUV_{peak}$ were also predictors of complete response. However, in general these indices did not allow differentiating NRs from PRs, a distinction that could be useful for patient management. For instance, within the patient population of our study the identification of PRs before any treatment could allow the definition of a subpopulation for which the use of conventional radiochemotherapy should be directly replaced by another option, such as a new targeted therapy.

A few studies have already focused on the link between image analysis and tumor biologic parameters. Gillies et al. (*33*) suggested that imaging can longitudinally characterize spatial variations in the tumor phenotype and its microenvironment so that the system dynamics over time can be quantitatively captured. Segal et al. (*22*) showed that contrast-enhanced CT image characteristics (such as texture heterogeneity score or estimated percentage of necrosis) correlate with most of the liver global gene expression profiles, revealing cell proliferation, liver synthetic function, and patient prognosis. Within the same context, Diehn et al. (*34*) mapped neuroimaging parameters with gene-expression patterns in glioblastoma, whereas Strauss et al. (*35*) combined dynamic PET kinetic parameters with gene-array techniques. Finally, Eary et al. (*24*) previously demonstrated that a globally assessed $^{18}$F-FDG distribution heterogeneity in sarcoma is a potential prognostic factor.

In our study, the value of textural feature analysis was explored on the pretreatment $^{18}$F-FDG PET scans for predicting response to combined chemoradiotherapy. Global tumor metabolic features based on the intensity histogram were computed directly on the original image. As such, they were therefore highly correlated with $^{18}$F-FDG uptake, which could explain why these textures could only predict CRs but could not distinguish NRs from PRs, similar to the SUV measurements. The other features evaluated in this study highlight tumor heterogeneity at a local and regional level, characterized in several ways, depending on the type of matrix used and the kind of feature computed on this matrix. Consequently, whereas a single feature cannot be directly linked to a specific biologic process, one could assume that a combination of textural parameters may be closely related to underlying physiologic processes, such as vascularization, perfusion, tumor aggressiveness, or hypoxia (*18,19*). Therefore textural features could be correlated to physiologic processes related to response to combined radiochemotherapy. For example, one could reasonably expect that a tumor exhibiting a heterogeneous, compared with a homogeneous, $^{18}$F-FDG distribution may respond less favorably to a uniformly distributed radiotherapy dose. We could also hypothesize that underlying neoangiogenesis contributes to tumor $^{18}$F-FDG uptake heterogeneity,

although it is now widely accepted that neoangiogenesis is associated with reduced effectiveness of conventional chemotherapy. However, the exact relationship between the proposed image-derived indices and underlying tumor biology can be established only on carefully designed prospective studies.

In this work, the cooccurrence features analyzing interrelationships between pairs of voxels, corresponding to the characterization of local nonuniformities, were able to significantly differentiate NRs from other patient groups. The measurement of local homogeneity and entropy gave the best results for this class of textures. Although in most cases responders (PR and CR) were associated with greater local heterogeneity than NRs, these features were less efficient in discriminating CRs from PRs.

The 2 features facilitating the best patient stratification were both associated with regional tumor characterization. Both the intensity and the size variability of uniform zones identified within the tumor, representing a measure of regional tumor heterogeneity, were significant predictors of response to therapy. ROC curve analysis showed that the performance of these features is similar to that of cooccurrence features in identifying NRs, but they can in addition distinguish between PRs and CRs with higher sensitivity and specificity than SUV measurements. These results suggest that regional (in terms of intensity and size of homogeneous areas) rather than local heterogeneity offers a superior differentiation of esophageal carcinoma patient groups in terms of response to combined chemoradiotherapy treatment than does any other global tumor metabolic activity measurement currently used in routine clinical practice, such as SUVs.

A limitation of the present study is that it is retrospective, considering a relatively small patient cohort. Therefore, the potential of new image-derived indices characterizing tumor $^{18}$F-FDG distribution for prediction of response to therapy studies demonstrated in this work needs to be validated by a prospective study on a larger patient cohort.

## CONCLUSION

We have demonstrated that textural analysis of the intratumor tracer uptake heterogeneity on baseline $^{18}$F-FDG PET scans can predict response to combined chemoradiation treatment in esophageal cancer. Textural features derived from cooccurrence matrices strongly differentiated NRs from PRs, providing useful information for personalizing patient management. These results suggest that regional and local characterization of $^{18}$F-FDG PET tracer heterogeneity in tumors, exploring processes underlying the $^{18}$F-FDG uptake and distribution within tumors, are more powerful than global measurements currently used in clinical practice, holding the potential to revolutionize the predictive role of PET in cancer treatment. Finally, although only $^{18}$F-FDG images in esophageal cancer have been considered here, clearly the same indices applied in other PET radiotracer studies in the same or different tumor types may help create even stronger links between imaging and underlying tumor biology.

## ACKNOWLEDGMENTS

This study was supported by a grant from the Ligue Contre le Cancer (Finistère and Côtes d'Armor Committees), IFR148-ScInBioS, and a fellowship from the French Ministry of Education and Research.

## REFERENCES

1. Gebski V, Burmeister B, Smithers BM, et al. Survival benefits from neoadjuvant chemoradiotherapy or chemotherapy in oesophageal carcinoma: a meta-analysis. *Lancet Oncol.* 2007;8:226–234.
2. Cunningham D, Allum WH, Stenning SP, et al. MAGIC trial participants: perioperative chemotherapy versus surgery alone for resectable gastroesophageal cancer. *N Engl J Med.* 2006;355:11–20.
3. Di Fabio F, Pinto C, Rojas Llimpe FL, et al. The predictive value of $^{18}$F-FDG PET early evaluation in patients with metastatic gastric adenocarcinoma treated with chemotherapy plus cetuximab. *Gastric Cancer.* 2007;10:221–227.
4. Krause BJ, Herrmann K, Wieder H, zum Büschenfelde CM. $^{18}$F-FDG PET and $^{18}$F-FDG PET/CT for assessing response to therapy in esophageal cancer. *J Nucl Med.* 2009;50:89S–96S.
5. van Westreenen H, Westerterp M, Bossuyt P, et al. Systematic review of the staging performance of $^{18}$f-fluorodeoxyglucose positron emission tomography in esophageal cancer. *J Clin Oncol.* 2004;22:3805–3812.
6. Ben-Haim S, Ell P. $^{18}$F-FDG PET and PET/CT in the evaluation of cancer treatment response. *J Nucl Med.* 2009;50:88–99.
7. Rizk N, Downey RJ, Akhurst T, et al. Preoperative 18FDG positron emission tomography standardized uptake values predict survival after esophageal adenocarcinoma resection. *Ann Thorac Surg.* 2006;81:1076–1081.
8. Javeri H, Xiao L, Rohren E, et al. Influence of the baseline 18FDG positron emission tomography results on survival and pathologic response in patients with gastroesophageal cancer undergoing chemoradiation. *Cancer.* 2009;115:624–630.
9. Rizk NP, Tang L, Adusumilli PS, et al. Predictive value of initial PET SUVmax in patients with locally advanced esophageal and gastroesophageal junction adenocarcinoma. *J Thorac Oncol.* 2009;4:875–879.
10. Flamen P, Van Cutsem E, Lerut A, et al. Positron emission tomography for assessment of the response to induction radiochemotherapy in locally advanced oesophageal cancer. *Ann Oncol.* 2002;13:361–368.
11. Downey RJ, Akhurst T, Ilson D, et al. Whole body 18FDG-PET and the response of esophageal cancer to induction therapy: results of a prospective trial. *J Clin Oncol.* 2003;21:428–432.
12. Kim MK, Ryu JS, Kim SB, et al. Value of complete metabolic response by $^{18}$F-fluorodeoxyglucose-positron emission tomography in oesophageal cancer for prediction of pathologic response and survival after preoperative chemoradiotherapy. *Eur J Cancer.* 2007;43:1385–1391.
13. Weber WA, Ott K, Becker K, et al. Prediction of response to preoperative chemotherapy in adenocarcinomas of the esophagogastric junction by metabolic imaging. *J Clin Oncol.* 2001;19:3058–3065.
14. Lordick F, Ott K, Krause BJ, et al. PET to assess early metabolic response and to guide treatment of adenocarcinoma of the oesophagogastric junction: the MUNICON phase II trial. *Lancet Oncol.* 2007;8:797–805.
15. Ott K, Weber WA, Lordick F, et al. Metabolic imaging predicts response, survival, and recurrence in adenocarcinomas of the esophagogastric junction. *J Clin Oncol.* 2006;24:4692–4698.
16. Cazaentre T, Morschhauser F, Vermandel M, et al. Pre-therapy $^{18}$F-FDG PET quantitative parameters help in predicting the response to radioimmunotherapy in non-Hodgkin lymphoma. *Eur J Nucl Med Mol Imaging.* 2010;37:494–504.
17. Vesselle H, Schmidt RA, Pugsley JM, et al. Lung cancer proliferation correlates with 18FDG uptake by positron emission tomography. *Clin Cancer Res.* 2000;6:3837–3844.
18. Rajendran JG, Schwartz DL, O'Sullivan J, et al. Tumour hypoxia imaging with $^{18}$F fluoromisonidazole positron emission tomography in head and neck cancer. *Clin Cancer Res.* 2006;12:5435–5441.
19. Kunkel M, Reichert TE, Benz P, et al. Overexpression of Glut-1 and increased glucose metabolism in tumours are associated with a poor prognosis in patients with oral squamous cell carcinoma. *Cancer.* 2003;97:1015–1024.

20. Xu Y, Sonka M, McLennan G, Guo J, Hoffman EA. MDCT-based 3-D texture classification of emphysema and early smoking related lung pathologies. *IEEE Trans Med Imaging.* 2006;25:464–475.
21. Tesar L, Shimizu A, Smutek D, Kobatake H, Shigeru N. Medical image analysis of 3D CT images based on extension of Haralick texture features. *Comput Med Imaging Graph.* 2008;32:513–520.
22. Segal E, Sirlin CB, Ooi C, et al. Decoding global gene expression programs in liver cancer by noninvasive imaging. *Nat Biotechnol.* 2007;25:675–680.
23. El Naqa I, Grigsby PW, Aptea A, et al. Exploring feature-based approaches in PET images for predicting cancer treatment outcomes. *Pattern Recognit.* 2009;42:1162–1171.
24. Eary JF, O'Sullivan F, O'Sullivan J, Conrad EU. Spatial heterogeneity in sarcoma $^{18}$F-FDG uptake as a predictor of patient outcome. *J Nucl Med.* 2008;49: 1973–1979.
25. Therasse P, Arbuck SG, Eisenhauer EA, et al. New guidelines to evaluate the response to treatment in solid tumors. European Organization for Research and Treatment of Cancer, National Cancer Institute of the United States, National Cancer Institute of Canada. *J Natl Cancer Inst.* 2000;92:205–216.
26. Hatt M, Cheze le Rest C, Turzo A, Roux C, Visvikis D. A fuzzy locally adaptive Bayesian segmentation approach for volume determination in PET. *IEEE Trans Med Imaging.* 2009;28:881–893.
27. Amadasun M, King R. Textural features corresponding to textural properties. *IEEE Trans Syst Man Cybern.* 1989;19:1264–1274.
28. Haralick RM, Shanmugam K, Dinstein I. Textural features for image classification. *IEEE Trans Syst Man Cybern.* 1973;3:610–621.
29. Loh H, Leu J, Luo R. The analysis of natural textures using run length features. *IEEE Trans Ind Electron.* 1988;35:323–328.
30. Thibault G, Fertil B, Navarro C, et al. Texture indexes and gray level size zone matrix: application to cell nuclei classification. *Pattern Recognition Inf Process.* 2009;140–145.
31. Hautzel H, Müller-Gärtner HW. Early changes in fluorine-18-FDG uptake during radiotherapy. *J Nucl Med.* 1997;38:1384–1386.
32. Erasmus JJ, Munden RF, Truong MT, et al. Preoperative chemoradiation-induced ulceration in patients with esophageal cancer: a confounding factor in tumor response assessment in integrated computed tomography-positron emission tomography imaging. *J Thorac Oncol.* 2006;1:478–486.
33. Gillies RJ, Anderson AR, Gatenby RA, Morse DL. The biology underlying molecular imaging in oncology: from genome to anatome and back again. *Clin Radiol.* 2010;65:517–521.
34. Diehn M, Nardini C, Wang DS, et al. Identification of noninvasive imaging surrogates for brain tumor gene-expression modules. *Proc Natl Acad Sci USA.* 2008;105:5213–5218.
35. Strauss LG, Pan L, Koczan D, et al. Fusion of positron emission tomography (PET) and gene array data: a new approach for the correlative analysis of molecular biological and clinical data. *IEEE Trans Med Imaging.* 2007;26: 804–812.

# Evaluation of a 3D local multiresolution algorithm for the correction of partial volume effects in positron emission tomography

Adrien Le Pogam[a)]
*PET Methodology, MRC Clinical Sciences Centre, Hammersmith Hospital Campus, Imperial College, W12 0NN London, United Kingdom*

Mathieu Hatt, Patrice Descourt, and Nicolas Boussion
*INSERM U650, LaTIM, CHU Morvan, 29609 Brest, France*

Charalampos Tsoumpas
*Division of Imaging Sciences, King's College, St. Thomas' Hospital, SE1 7EH London, United Kingdom*

Federico E. Turkheimer
*PET Methodology, MRC Clinical Sciences Centre, Hammersmith Hospital Campus, Imperial College, W12 0NN London*

Caroline Prunier-Aesch, Jean-Louis Baulieu, and Denis Guilloteau
*INSERM U930, Nuclear Medicine Department, François Rabelais University and University-Hospital Bretonneau, 37044 Tours, France*

Dimitris Visvikis
*INSERM U650, LaTIM, CHRU Morvan, 29609 Brest, France*

(Received 18 August 2010; revised 31 May 2011; accepted for publication 16 June 2011; published 9 August 2011)

**Purpose:** Partial volume effects (PVEs) are consequences of the limited spatial resolution in emission tomography leading to underestimation of uptake in tissues of size similar to the point spread function (PSF) of the scanner as well as activity spillover between adjacent structures. Among PVE correction methodologies, a voxel-wise mutual multiresolution analysis (MMA) was recently introduced. MMA is based on the extraction and transformation of high resolution details from an anatomical image (MR/CT) and their subsequent incorporation into a low-resolution PET image using wavelet decompositions. Although this method allows creating PVE corrected images, it is based on a 2D global correlation model, which may introduce artifacts in regions where no significant correlation exists between anatomical and functional details.
**Methods:** A new model was designed to overcome these two issues (2D only and global correlation) using a 3D wavelet decomposition process combined with a local analysis. The algorithm was evaluated on synthetic, simulated and patient images, and its performance was compared to the original approach as well as the geometric transfer matrix (GTM) method.
**Results:** Quantitative performance was similar to the 2D global model and GTM in correlated cases. In cases where mismatches between anatomical and functional information were present, the new model outperformed the 2D global approach, avoiding artifacts and significantly improving quality of the corrected images and their quantitative accuracy.
**Conclusions:** A new 3D local model was proposed for a voxel-wise PVE correction based on the original mutual multiresolution analysis approach. Its evaluation demonstrated an improved and more robust qualitative and quantitative accuracy compared to the original MMA methodology, particularly in the absence of full correlation between anatomical and functional information.  [DOI: 10.1118/1.3608907]

Key words: emission tomography, partial volume effects, resolution and intensity recovery, wavelet transform, multimodality

## I. INTRODUCTION

Partial volume effects (PVEs) refer to two distinct phenomena leading in underestimation or overestimation of the tissues uptake. The first results from the limited spatial resolution of the imaging device, leading to a three-dimensional (3D) blurring and a loss of signal in tissues of size similar to the system's point spread function (PSF) full width at half-maximum (FWHM), as well as activity cross-contamination ("spillover") between structures with different uptakes.[1] The second phenomenon arises from the discrete representation on a grid with voxel sizes from 1 to 5 mm for the reconstruction of images. The voxel values at the edges are consequently a mixture of different tissues, an effect

commonly known as "tissue-fraction effect." This effect exists on all image modalities, but its magnitude is lower on anatomical datasets such as magnetic resonance image (MRI) or computed tomography (CT), and the introduction of higher resolution details from anatomical images in the functional images could reduce its impact. In this study, we considered both the spatial resolution of the scanner and the voxel grid sampling, which are specific to the low-resolution images obtained with PET.

Most of the previously proposed approaches for PVE correction consist in using *a priori* anatomical information provided by either computed tomography or magnetic resonance images. Although these methodologies usually aim at the recovery of accurate uptakes in specific regions of interest (ROIs)[1–4] requiring either a coregistered atlas or a segmentation step, such as the ROI based geometric transfer matrix (GTM) by Rousset *et al.*,[2] some voxel-based implementations have been recently proposed. They could be classified in two groups; namely deconvolution without the use of anatomical information[5–8] or voxel-wise correction based on the use of anatomical and functional images available from multimodality devices.[9–11]

Teo *et al.*[5] considered the use of iterative deconvolution restricted to a ROI to avoid noise increase in the overall image, while a more recent study applied iterative deconvolution to brain images.[6] Boussion *et al.*[7] proposed the use of wavelet-based denoising at the iterative level of the deconvolution process to avoid noise amplification. The efficacy of this denoising greatly depends on the choice of the wavelet filtering algorithm and associated threshold values. Kirov *et al.*[8] suggested the use of regularized iterative deconvolution with variance control based on local topology as an alternative solution for noise reduction. This latter approach is not fully automatic because the determination and optimization of the regularization parameters are dependent on the properties of the image.

One methodology based on the use of anatomical information to correct functional data was proposed by Boussion *et al.*[9] and is referred to as mutual multiresolution analysis (MMA). It uses discrete wavelet transforms[10] of a low-resolution PET image L and a corresponding coregistered high resolution anatomical MRI (or CT) H. The method extracts the spatial frequencies like details, edges, and textures from wavelet decompositions at a level of resolution common to H and L (a specific decomposition layer in which both wavelet images have the same PSF FWHM). A global linear model is then inferred to build the lacking details of L from these found in H. The method demonstrated accurate quantitative correction comparable with the methodologies considered as the current state of the art[2–4] but limited to ROI analyses. This approach has the advantage of generating PVE corrected images, allowing for an accurate activity recovery, without any segmentation or other preprocessing steps.

Despite these advantages, the original MMA approach suffers from two limitations. First, it is based on a global correlation between anatomical and functional structures, as a linear and global link is assumed in order to model the relation between the wavelet coefficients of both modality transforms at the same level of resolution. Consequently, where there is little to no correlation between these structures, artifacts may appear in the corrected emission image as a result of the incorporation of anatomical structures with no functional significance. An alternative approach based on MMA has been recently published[11] restricted to the brain domain and making use of atlases for the provision of the anatomical information in an attempt to reduce the impact of functional and anatomical image mismatches. However, such approach is dependent on the use of atlases and is, therefore, only applicable to brain imaging. The second limitation of the original MMA algorithm is due to the use of a 2D modeling as the correction is applied independently slice by slice, whereas PVE is a 3D effect for which it is important to consider interactions in all three spatial directions.

The goal of the present study was to design a new model in order to overcome the shortcomings associated with the original MMA algorithm as highlighted above. More specifically, a new 3D wavelet decomposition scheme was designed, and the global linear relation model was replaced with an improved local analysis, in order to process limited image areas at a time and adapt this model to each image part based on local information. This new approach was evaluated on synthetic and simulated images and applied to brain and whole body patient datasets, and its performance was compared to the results obtained with both the original 2D global MMA algorithm described by Boussion *et al.*[9] and the GTM reference methodology.[2]

## II. MATERIALS AND METHODS

### II.A. Multiresolution analysis in the wavelet domain

Performing multiresolution image analysis using wavelet transforms consists in analyzing details across different levels of resolution or scale,[12] in order to extract consecutive layers of details from large structures to small edges, by separating the spatial frequencies they contain. Among the many algorithms developed to perform wavelet transform of an image, the most common approach is the multiresolution pyramidal methodology[13] consisting in iteratively reducing the resolution of the image. Such subsampling may cause a loss of linear continuity in spatial features such as edges and the appearance of artifacts in those structures.[14] Therefore, the undecimated algorithms are often more appropriate, for instance, for the image fusion purpose.[15] The "à trous" ("with holes") algorithm was used here and extended to 3D to perform the wavelet decomposition[10] based on the initial work of Boussion *et al.*[9] This algorithm presents several advantages such as a straightforward implementation, a reconstruction without any loss of information and an isotropic process (i.e., no specific directions selected). This algorithm is related to the standard discrete wavelet transform decomposition scheme[16] as it performs the subsampling of the filtered image by upsampling the low-pass filter, inserting zeros between each of the filter's coefficient at each level. The detail coefficients images $\{w_j\}$ are then obtained as the difference $\{I_{j+1} - I_j\}$

between the low-pass filtered images from two consecutive levels. At each iteration j, the resolution of the image $I_j$ is reduced to obtain the smoothed image $I_{j+1}$ (called residual) using

$$I_{j+1}(x,y,z) = \sum_{m,n,o\in[-2;2]} h(m,n,o) I_j (x+m.2^j, y+n.2^j, z+o.2^j), \quad (1)$$

where h is a 3D low-pass filter defined by

$$\begin{cases} h(x,y,z) = h_{1D}(x) * h_{1D}(y) * h_{1D}(z) \\ h_{1D}(0) = 3/8, h_{1D}(\pm 1) = 1/4, h_{1D}(\pm 2) = 1/16, \text{ and } h_{1D}(n) = 0 \text{ if } |n| > 2 \end{cases}, \quad (2)$$

with $*$ the convolution operator and $h_{1D}$ a binomial filter[17] of order 4. The inverse transform can be computed by adding the detail layers $\{w_j\}$ from all levels to the final low-resolution image $I_J$

$$I_0(x,y,z) = I_J(x,y,z) + \sum_{j=1}^{j=J} w_j(x,y,z), \quad (3)$$

where J is the number of iterations from the initial image $I_0$ to the final approximation $I_J$.

Due to the discrete convolution shown in Eq. (1), the spatial resolution (or PSF FWHM) of the residual image $I_{j+1}$ depends on both the spatial resolution of $I_j$ and the size of its voxels. This is because the image $I_j$ is sampled according to the iteration index j (i.e., the level of scale in the wavelet transform) and the convolution leads to a nonlinear dependency (factor $2^j$) on the resulting PSF FWHM of the residual $I_{j+1}$. The relationship between voxel sizes and the residual FWHM at different scales was determined by applying the à trous algorithm to an initial point source of different sizes (from 0.6 to 1.4 mm) using the MATHEMATICA software. This relationship was found to be linear (see Table I). In order to change the voxel size of the initial image $I_0$ and to obtain a residual $I_j$ with a specific spatial resolution according to the MATHEMATICA software analysis, trilinear interpolation and third order B-spline resampling were used for anatomical and functional images, respectively.

Using both Eq. (3) and this resampling, the spatial frequencies $\{w_j\}$ of the anatomical and functional images are extracted in 3D and the spatial resolution of the final residual $I_J$ is accurately determined. We define the spatial resolution of the anatomical image H as q (initial image H referred to as $H_q$) and that of the functional image L as $r = q + p$ (initial image L referred to as $L_r$), with $r > q$ and p the number of decompositions that have to be performed to reach a common level of resolution between the two modalities. This number is deduced from the PSF FWHM and the voxel sizes of both initial images. We can then perform the extraction of the spatial frequencies at a level of resolution common to H and L $(q+p+1)$, using the à trous algorithm

$$L_r(x,y,z) = L_{q+p}(x,y,z) = L_{q+p+1}(x,y,z) + w^L_{q+p+1}(x,y,z), \quad (4)$$

$$H_q(x,y,z) = H_{q+p+1}(x,y,z) + \sum_{k=1}^{k=p+1} w^H_{q+k}(x,y,z), \quad (5)$$

where $w^L$ (in terms of units) and $w^H$ are the wavelet coefficients from the wavelet decomposition of the functional and anatomical images, respectively.

TABLE I. FWHM values of the residual scales for an initial FWHM of 1 mm and initial voxel sizes of 0.6, 1, and 1.4 mm. The equation formula provided by the software is presented.

| Voxel size of the initial image (mm) | 0.60 | 1.00 | 1.40 | x |
|---|---|---|---|---|
| FWHM residual scale 1 | 1.56 | 2.61 | 3.65 | 2.61 * x |
| FWHM residual scale 2 | 3.38 | 5.64 | 7.90 | 5.64 * x |
| FWHM residual scale 3 | 6.89 | 11.48 | 16.08 | 11.48 * x |

### II.B. Local mutual multiresolution analysis

Similarly to previously developed methodologies using wavelets for PVE correction,[9,11] our method is based on the exploitation of existing correlations in the wavelet domain. Such correlation is defined by establishing a relationship between the anatomical and functional wavelet coefficients at a common level of resolution, independently of the original images content (in terms of units). In our proposed approach, we assume that the residual $L_r$ can be estimated by a space dependent scaling of $H_r$, for which the scaling factor can be obtained from a local analysis to account for local differences between H and L. The corrected emission image Lq (same level of resolution as Hq) is, therefore, obtained using Eq. (6) by adding the original uncorrected value of L $[L_{q+p+1}(x,y,z) + w^L_{q+p+1}(x,y,z)$, see Eq. (4)] with the sum of the anatomical wavelet coefficients weighted by a local factor $\alpha$

$$L_q(x,y,z) = L_{q+p+1}(x,y,z) + w^L_{q+p+1}(x,y,z) + \alpha(x,y,z) \sum_{k=1}^{k=p} w^H_{q+k}(x,y,z), \quad (6)$$

where $\alpha$ stands for the median of the ratio map (MRM). Equation (6) provides the corrected value for each voxel (x, y, z) by adding to the original functional image value (L) the wavelet coefficient. Each voxel is processed individually, however, the parameter $\alpha$ is computed for each voxel using a

3D sliding cube, simultaneously applied within the L and H wavelet layers to obtain the MRM

$$\alpha(x,y,z) = \text{median} \times \left\{ \frac{w^{L}_{q+p+1}(x_i,y_i,z_i)}{T(w^{H}_{q+p+1}(x_i,y_i,z_i))}, (x_i,y_i,z_i) \in \text{WIND} \right\}, \quad (7)$$

where WIND is a window in 3D (i.e., a cube) centered on (x,y,z) with a fixed size of $3 \times 3 \times 3$ voxels ($i = 1 \ldots 3$) and T is a threshold operator. Equation (7), therefore, computes $\alpha$ for each voxel of coordinates x, y, z, as the median (within a cube centered on this voxel) of the ratios between the wavelets coefficients of each transforms. The cubic moving window was introduced instead of a simple voxel-by-voxel ratio to account for local variations in the anatomical and functional images, contrary to the global model[9] and to reduce the noise and misregistration sensitivity associated with a voxel-by-voxel analysis. The choice of using the median instead of the mean within the cubic sliding window, as well as the actual size of the cube was made based on the results obtained on simulations (see Sec. IV). Different sizes (3, 5, and 7 voxels) were compared since this parameter might have an impact on the sensitivity of the method to spatial misregistration, noise and inappropriate choice of the respective PSF FWHM values. Also, results obtained using the mean or the median of the ratios were compared. The mean was expected to be more sensitive to artifacts and noise in contrast to the median, which would tend to discard extreme values. Also, the use of a sliding cube should also make the approach more robust to spatial misregistration or inaccurate PSF FWHM values used in the process. On the one hand, the ratio map allows a large amount of uncorrelated details such as anatomical structures without significant uptake in the functional image to be discarded. On the other hand, if no structural information is associated with a significant uptake in the corresponding functional image, the low values of the wavelet coefficients may lead to MRM evaluation errors with denominator values $w^{H}_{q+p+1}$ close to zero. In order to avoid extreme values that may be generated by the MRM and preserve the activity of the functional image in such configurations, we introduced a fixed threshold T [Eq. (7)] on the anatomical wavelet coefficients. Its value was empirically chosen as 0.1 as it gave satisfactory results in most considered cases, however future studies should investigate the automatic estimation of an optimal value for each case. Therefore, an investigation regarding the cube optimal size was carried out in this study.

## II.C. Validation and comparison study

### II.C.1. Analysis

Mean and associated standard deviation were computed on ROI placed on the organs or objects of interest in order to quantify the partial volume effect correction and compared to the ground-truth when available (synthetic images and simulated datasets). In such cases (for example, the spheres), the ROIs were defined on the ground-truth, covering the entire structure voxel-by-boxel, and were reported to both uncorrected and corrected images. For clinical datasets, as ground-truth was not available, the improvement between uncorrected and corrected image was reported for ROI placed on both images.

### II.C.2. Synthetic images

The proposed algorithm was first validated using synthetic images. All the images considered in this section were generated in 3D and analyzed either on a 2D basis using the original MMA method, or in 3D using the proposed approach. Finally, the GTM methodology was also applied, assuming a perfect knowledge of the ground-truth for the definition of the necessary ROIs in order to eliminate any potential errors that can be associated with a segmentation step. Two different synthetic images were employed, the first one $H_{ref}$ [Fig. 1-(1)] to generate the functional images L [Fig. 1-(2)], and the second one $H_{anat}$ [Fig. 1-(3)] used for the correction of PVE via both the 2D global and 3D local methodologies. A Gaussian noise [standard deviation (SD) 2% of the mean in the uniform part of the phantom] was added in the $H_{anat}$ images prior to their use for the correction. These $128 \times 128 \times 128$ ($1 \times 1 \times 1$ mm$^3$ voxels) images contain a cylindrical background region with a fixed intensity of 100 and spheres of different sizes and intensities. The first [Fig. 1(a)-(1)] contains five 2 cm diameter spheres with decreasing intensities (200, 120, 90, 70, and 50). The second one [Fig. 1(b)-(1)] contains four spheres of decreasing diameter (6, 4, 2, and 1 cm) with intensity of 200. The last two images [Figs. 1(c)-(1) and 1(c)-(3)] display spheres common to both modalities with, however, no absolute intensity correlation, as well as two additional structures: one which is present only in the anatomical data $H_{anat}$ with no corresponding uptake in the functional image and, respectively, a hot spot in the functional data L without any corresponding anatomical structure in $H_{anat}$.

The L images [Fig. 1-(2)] (same voxel size and dimensions as for $H_{anat}$) were generated by convolving $H_{ref}$ [Fig. 1-(1)] with a 6 mm FWHM 3D Gaussian PSF and adding Gaussian noise (SD 10% of the mean intensity). In this first dataset, both $H_{anat}$ and L images had a voxel size of 1 mm$^3$. The Gaussian noise approximation is realistic enough for the reconstructed PET images when considering a specific ROI.[18] The noise intensity (SD value) used was determined through different ROI analyses in the lung and liver from various whole body clinical datasets.

Different combinations of the 3D synthetic images in Fig. 1 were considered to compare the performance of the different approaches considered. First, functional and anatomical images with complete structural and intensity correlation [Figs. 1(a)-(2) and 1(a)-(3)] were used to specifically study the accuracy of the correction for spillover effects due to the various contrasts. A second combination was analyzed [Figs. 1(b)-(2) and 1(b)-(3)] in order to examine the value recovery of small objects. These two configurations were designed to validate the performance of the local approach for cases where the global approach already leads to satisfactory results, i.e., with a perfect match (structure and intensity)

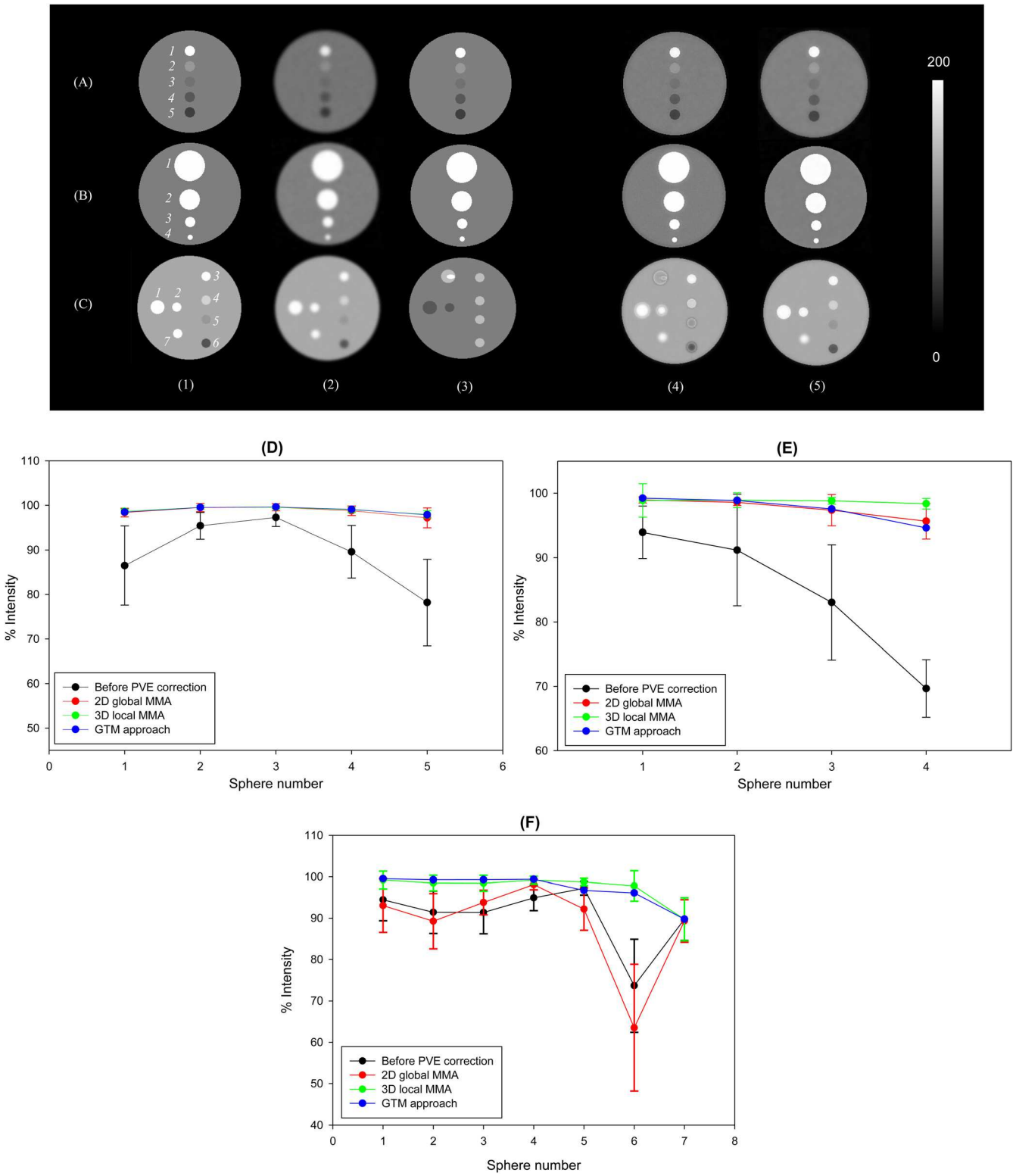



FIG. 1. Three synthetic datasets for which the background region has a fixed value of 100. (a) A correlated case with spheres of constant size and decreasing intensities (from 200 to 50); (b) a correlated case with spheres of constant intensity (200) and decreasing sizes; and (c) an uncorrelated case. For each one of the three synthetic datasets: (1) a high-resolution anatomical image $H_{ref}$ used to generate the low-resolution functional image L (2), (3) the high-resolution anatomical image $H_{anat}$ used for the PVE correction, (4) the PVE corrected images using the 2D global approach, and (5) images after PVE correction using the 3D local approach. Percentage of intensity recovery (mean ± SD) for the three datasets corrected for PVE using the GTM (mean only), 2D global or 3D local approach: (d) a correlated case with spheres of constant size and decreasing intensities, (e) a correlated case with spheres of constant intensity and decreasing sizes, and (f) an uncorrelated case. Spheres are numbered on 1-(1).

between both modalities. A third combination was used to investigate partial correlation between the two modalities [Figs. 1(c) (2), and (3)] for which the 2D global MMA is expected to perform with reduced qualitative and quantitative accuracy. In each case, the new local 3D MMA approach was compared to the previous 2D global model and to the GTM method.

Finally, tests similar to those already used for the 2D MMA (Ref. 9) were carried out to evaluate the impact of noise and possible misregistration errors, or an inappropriate choice of FWHM parameters. For this purpose, variable noise intensities (SD from 10% to 50%) were added to the functional images of Fig. 1(c)-(2). The robustness of the approach against misregistration errors was evaluated by applying different rigid transformations [translation up to 4 voxels (about 4 mm), rotation up to 5°] or inappropriate scaling (±10%) to the functional image of Fig. 1(c)-(2). The impact of an inappropriate choice of FWHM parameters for the resampling was investigated by generating different L images from the $H_{ref}$ [Fig. 1(c)-(1)] with FWHM PSF from 4 to 8 mm. These L images were then corrected for PVE considering a unique FWHM value of 6 mm.

### II.C.3. Simulated images

In the second part of this study, simulated images generated using a segmented brain phantom based on measured T1 MRIs (Ref. 19) were analyzed. The images are a T1-weighted MRI and an associated $^{18}$F-FDG PET. The functional image was generated using the following procedure.[20] Clinically measured plasma time activity curves (TACs) were first used to generate a set of TACs for each anatomical region of the brain phantom according to 28 different clinical dynamic frames ($1 \times 30$ s, $1 \times 15$ s, $1 \times 5$ s, $4 \times 10$ s, $4 \times 30$ s, $4 \times 60$ s, $4 \times 120$ s, $9 \times 300$ s). Pathological parameters were introduced in the parietal and the anterior frontal lobes. The dynamic images were then forward-projected using the maximum ring difference, mash and span as for the patient study, forming projection data of the ECAT HR+ scanner (spatial resolution of 4.8 and 5.6 mm FWHM transaxially and axially, respectively). These projection data were attenuated using the values of the different tissue types contained in the Zubal phantom (muscle, bone, fat, and skin), and uncorrected for normalization by applying the inverse normalization factors. These factors and the scatter additive term were both taken from the human study. The random contribution was simulated based on the system's detection efficiency factors pattern scaled to the total random events of each frame in the acquired human study. Finally, Poisson noise was added to the sinograms, and the images were finally reconstructed with filtered backprojection including scatter, attenuation, and normalization corrections. The image sizes for both PET and MRI were $128 \times 128 \times 64$ ($2.35 \times 2.35 \times 2.42$ mm$^3$ voxels). The final static PET image was eventually obtained by summing the last six temporal frames. The respective considered FWHM values were 4.8 mm in plane and 5.6 mm axially for the PET images and 1 mm in all three dimensions for the anatomical datasets. The performance of the method was assessed by comparison with the known ground-truth of the simulation. This dataset was also used to determine the optimal sliding cube size among $3 \times 3 \times 3$, $5 \times 5 \times 5$ or $7 \times 7 \times 7$ voxels (see Sec. IV).

### II.C.4. Clinical images

The approaches were also compared on two clinical images. The first one consists of a clinical T1 MRI (GE 1.5T) and associated FDG brain PET (Philips GEMINI dual slice PET/CT) scans. The MRIs were chosen instead of the CT for the correction to benefit from the improved contrast in the different brain structures. The MRI contains a hyperintensity signal in the left occipital lobe and the posterior cingulum due to the gadolinium injection. The PET reconstructed images [using RAMLA 3D (two iterations, relaxation parameter of 0.05 and a 5 mm FWHM 3D Gaussian postfiltering] and CT based attenuation correction) are $128 \times 128 \times 64$ ($1.41 \times 1.41 \times 2$ mm$^3$ voxels), and the MRI is $512 \times 512 \times 160$ ($0.47 \times 0.47 \times 1$ mm$^3$ voxels). The PET image and MRI were spatially coregistered using mutual information maximization and affine transformations using MIPAV software [Center for Information Technology (CIT) National Institutes of Health (NIH)]. The FWHM considered was 1 and 5 mm in all three dimensions for the MRI and PET datasets, respectively (considering the spatial resolution for the Philips GEMINI PET system of 5.2 and 5.4 mm FWHM transaxially and axially, respectively, in combination with the 5 mm FWHM Gaussian postfiltering applied to the reconstructed image). Qualitative evaluation was carried out using profiles through the frontal and the tempo-occipital regions. Quantitative accuracy was evaluated by white and gray matter quantification (mean intensity and standard deviation) before and after PVE correction using an automatic delineation on the MRI via the voxel-based morphometry segmentation tool of the SPM software.[21] The same delineation results were used for the ROI based correction using GTM for comparison. The impact of the 3D analysis was observed on this dataset in which voxels are anisotropic.

The second dataset is a whole-body $^{18}$F-FDG-PET/CT image of a lung cancer patient (GE Discovery STE 4-slice PET/CT), acquired 55 min after injection of 355 MBq (CT: 80 mA, 140 kVp, PET: 3 min per axial field of view). PET images were reconstructed (voxel size of $4.68 \times 4.68 \times 3.27$ mm$^3$ and a matrix size of $128 \times 128 \times 47$ voxels) using OSEM (two iterations, 28 subsets) and CT based attenuation correction. The resolution of the original CT image was $0.97 \times 0.97 \times 0.97$ mm$^3$ voxels (matrix dimension $421 \times 321 \times 100$). The FWHM of the PET image was considered as 6.1 and 6.7 mm in plane and axially, respectively (matching the spatial resolution of GE Discovery STE), and 1 mm in all three directions for the CT images. Manually drawn spherical regions were placed on the lesion ($ROI_{lesion}$) and in the lung ($ROI_{normal}$). An additional ROI in the spinal region ($ROI_{bone}$) was used in order to evaluate the potential of introducing artifacts in the PVE corrected PET images as a result of prominent anatomical structures (such as bones) in the CT images, which are not present in the FDG PET images. Since

no ground-truth is available in this case, anatomical images were semiautomatically segmented to generate ROIs for the tumor, the lungs, the soft tissues, and the bones in order to be used in the GTM method.

## III. RESULTS

### III.A. Synthetic images

The quantitative results regarding the correction using the three methods can be found in Figs. 1(d)–1(f) for the three different synthetic datasets. Figure 1(d) contains results of the spillover effects correction for the spheres of constant size and varying intensities [Fig. 1(a)]. It displays the percentage of recovered intensity (a value of 100 is a perfect recovery of the true activity) in the spheres, demonstrating similar levels of recovery for the three methods, within 2% for all of the different intensities considered. Figure 1(e) shows quantification results for the constant intensity and variable size spheres [Fig. 1(b)]. The recovered intensities in the spheres demonstrate that both approaches perform with similar accuracy. As the functional image is resampled with the voxel size of the anatomical image as a first step for both methodologies, the impact of tissue-fraction effect is reduced by introducing the higher resolution details of the anatomic imaging when available and correlated, achieving accurate intensity recovery even for the smallest spheres where PVE have the most significant impact. Figures 1(c)-(4) and (5) show the results for the uncorrelated case in which the spheres are different in the functional and anatomical images in terms of both intensity and structure, revealing differences between the 2D global and new 3D local approaches as the new approach does not incorporate uncorrelated anatomical details in the functional image during the correction, whereas the 2D global MMA creates local artifacts related to existing mismatches. In addition, the local approach handles more accurately the differences in absolute signal intensity between the anatomical and functional images. A quantitative comparison of recovered intensities in the spheres is shown in Fig. 1(f) and demonstrates much higher accuracy for the new approach: without correction, the mean error was $-11.4\% \pm 5.2\%$, whereas the 2D global and 3D local corrections resulted in a mean error of $-15.1\% \pm 6.1\%$ and $2.8\% \pm 2.4\%$, respectively. The GTM method resulted in a mean error of 3.2%. The standard deviation of each measurement associated with the use of the 2D global MAA approach was much larger than when applying the new model. This is explained by the artifacts that are introduced due to mismatch in structural information between anatomical and functional datasets that the 2D global approach is not able to address. Such a case highlights the limitation of the 2D global MMA and the way the new 3D approach successfully overcomes such issues, with similar accuracy to GTM using the ground-truth ROIs. The measurements in sphere 7 illustrate the fact that no PVE correction can be performed due to the lack of corresponding structure in the anatomical image. It is important, however, to emphasize that in such a case, the functional activity in the corrected image remains unchanged and no additional artifact is introduced.

The local approach appeared more robust with respect to the increase of noise in the functional image in Fig. 1(c)-(2): mean recovered intensity across all structures (excluding sphere 7) and the different noise levels was $99.5\% \pm 1.1\%$ for the 3D local, whereas the 2D global results exhibited much higher standard deviation with a mean recovery intensity of $97.8\% \pm 27.8\%$ as illustrated in Fig. 2(a).

The impact of an inappropriate choice of FWHM value and of a spatial misalignment between the anatomical and functional images was investigated for the 3D local MMA only as such an analysis has already been performed by Boussion *et al.*[9] for the 2D global approach only. This analysis demonstrated overall satisfactory robustness with, however, recovery errors reaching 25%–50% for certain spheres although the spatial misalignments and rotations considered were smaller (up to 3 mm and 3°, respectively). For the present study of the 3D local MMA robustness, a maximum mean error of $1.9\% \pm 23.1\%$ was reached with the investigated misalignment, rotations, and inadequate scaling parameters [Fig. 2(b)]. Regarding the inadequate choice of FHWM parameters, the mean error was $9.4\% \pm 5.5\%$ and $2.8\% \pm 3\%$ for $\pm 2$ mm and $\pm 1$ mm around the actual exact value [Fig. 2(c)].

### III.B. Simulated images

Figures 3(b) and 3(c) illustrate the correction obtained using the mean and the median of the ratios, respectively (use to establish the factor $\alpha$), on the simulated $^{18}$F-FDG brain PET and associated T1-weighted MRIs of Fig. 3(a). The impact of the cube size ($3 \times 3 \times 3$, $5 \times 5 \times 5$ and $7 \times 7 \times 7$ voxels in (1), (2), and (3), respectively) is also illustrated. Irrespectively of the cube size, the mean of the ratio led to major artifacts due to noise and the possible approximations in the FWHM parameters. On the contrary, the use of the median led to better visual result. Only small visual differences (mostly a slight increasing blurring effect) were observed with the three different cube sizes [Figs. 3(c)-(1) to 3(c)-(3)]. This blurring effect can be explained by the inclusion of additional voxels for the computation of the median value and a less efficient local modeling. The quantitative analysis as shown in Fig. 3(d) demonstrated higher activity recovery when using the smallest cubic window size ($3 \times 3 \times 3$). All other results were, therefore, generated with this setting.

Figure 4(a) shows the results obtained on the simulated $^{18}$F-FDG brain PET images. The global 2D MMA led to the incorporation of all the MRI details into the corrected PET images [Fig. 4(a)-(3)], creating artifacts such as uptake corresponding to the skull, whereas the image corrected with the new approach [Fig. 4(a)-(4)] was free of such artifacts. Further evaluation using a frontal region profile is presented in Fig. 4(b), demonstrating higher contrast with the local approach. Both gray and white matters are better delineated. The spikes on both sides of the profile [see red arrows in Figs. 4(a)-(3) and 4(b)], corresponding to the artifact uptake from the bone incorporation using the global MMA

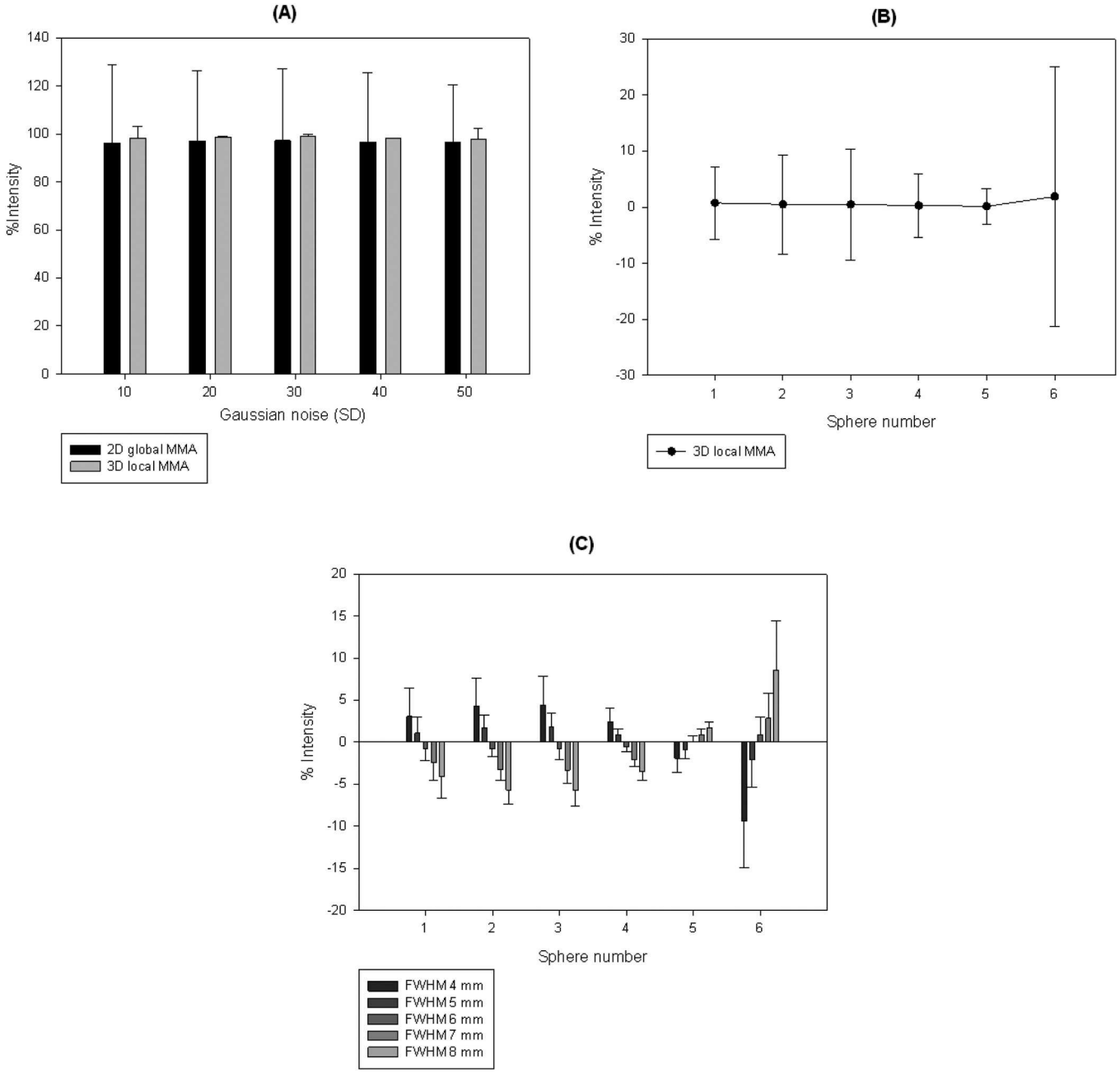



FIG. 2. (a) Percentage of mean recovered intensity relative to true values (mean ± SD) across all structures of Fig. 1(c)-(1) (excluding sphere 7) and the different noise levels (SD from 10% to 50%) in Fig. 1(c)-(2). Percentage error of recovered intensity relative to true values (mean ± SD) in the different spheres [Fig. 1(c)-(1)] considering: (b) different spatial misalignment scenarios between Figs. 1(c)-(2) and 1(c)-(3) (translation up to 4 voxels, rotation up to 5°) or inappropriate scaling (± 10%) of Fig. 1(c)-(2); (c) different PSF sizes (4–8 mm) in the PVE correction process relative to the true value (6 mm).

approach, are absent from the corrected image using the new model. Comparison of ROI values placed in different regions of the brain against the true image [Fig. 4(c) (1–5)] demonstrated higher accuracy of the correction with the new approach. In some regions such as amygdala, cerebellum, or thalamus, the 2D global MMA and the GTM approach accurately corrected the intensities. However, in other regions such as the frontal or the hippocampus region, for instance, it led to an overestimation or underestimation of the uptake, respectively. Mean error for all analyzed regions was 31.9% ± 8.5%, 21.3% ± 6.8%, 16.7% ± 5.7%, and 8.9% ± 2.7% for the noncorrected PET, 2D global MMA, GTM, and 3D local MMA, respectively.

### III.C. Clinical images

As a last evaluation step, the new approach was applied and compared to the global MMA and GTM on two clinical cases: a brain and a whole-body acquisition (Figs. 5 and 6, respectively). Despite, the good correlation between the T1 MRI [Fig. 5(a)-(1)] and FDG PET [Fig. 5(a)-(2)] images regarding the gray and white matter, noncorrelated

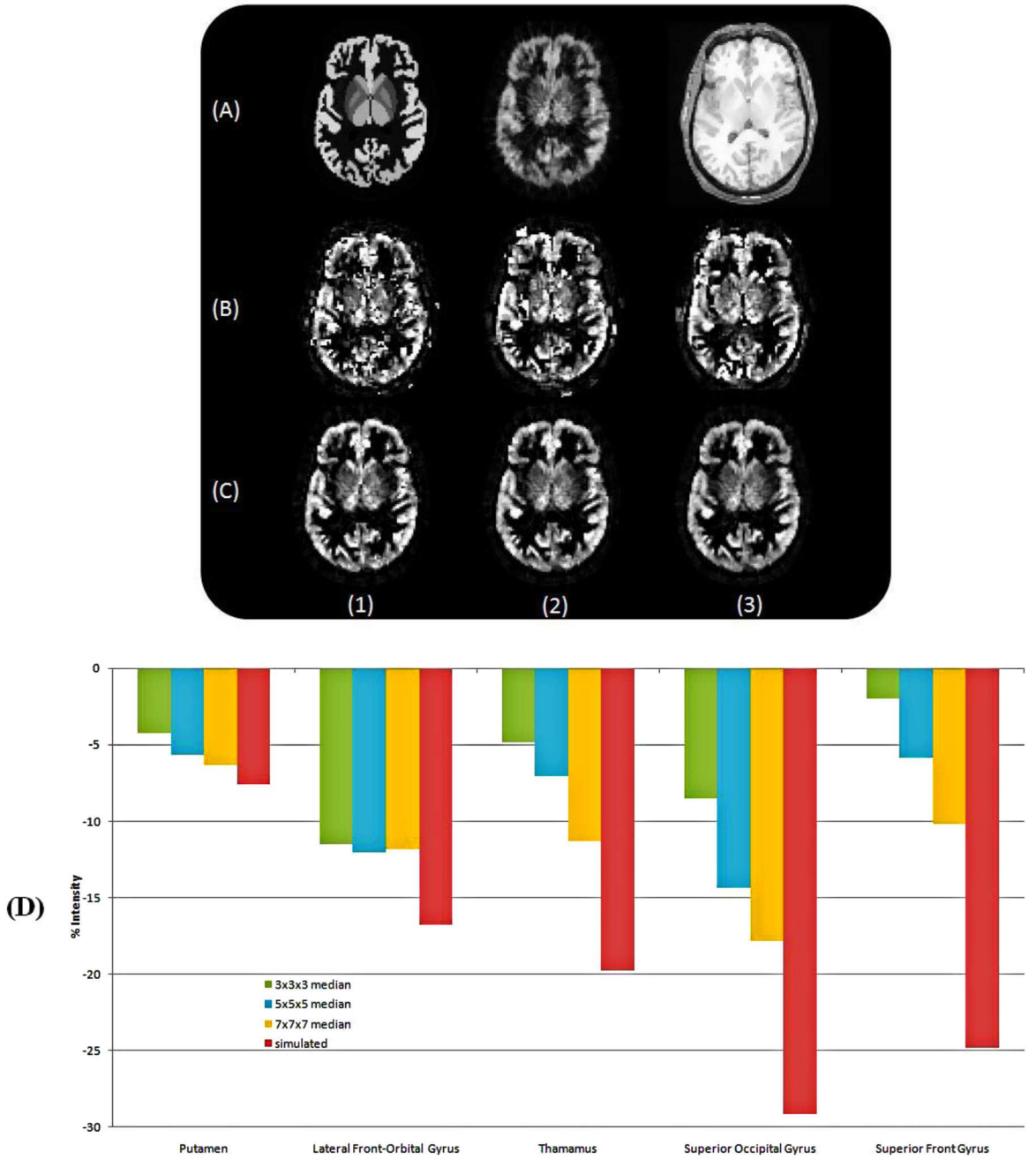



FIG. 3. Optimization and validation for the local modeling parameters on (a) a simulated FDG PET (a)-(2)/MRI (a)-(3) brain dataset [ground-truth on Fig. (a)-(1)] using; (b) the mean; and (c) the median of the ratio maps methodologies for, respectively a (1) $3 \times 3 \times 3$, (2) $5 \times 5 \times 5$, and (3) $7 \times 7 \times 7$ 3D cube. (d) Percentage intensity recovery in the different brain structures considered in the simulation following the PVE correction.

structures also exist, such as the skull (see red arrows) and the gadolinium enhancement (see white circle) in the MRI. These MRI features, which do not match any FDG uptake, were introduced in the corrected image by the 2D global process [Fig. 5(a)-(3)], whereas the 3D local approach [Fig. 5(a)-(4)] suppressed these uncorrelated details, leading to a more reliable and visually satisfying PVE correction. In addition, the new approach led to higher contrast between regions as shown in the profiles of Fig. 5(b). Furthermore, using the SPM software to segment the gray and white matter from the MRIs, we obtained quantitative results for the comparison of the two voxel-wise correction methodologies as well as corrected values using the GTM approach in these ROIs. The 2D global MMA led to higher standard deviation values due to the incorporation of additional uncorrelated details, with an average mean intensity variation with respect to the initial PET image in the gray and white matter regions of $1.8\% \pm 21.1\%$ and $0.2\% \pm 31.6\%$, respectively. In contrast, higher mean intensity variations and lower standard deviation were obtained with the new approach: $11.4\% \pm 6.5\%$ in the gray matter and $-2.6\% \pm 8.3\%$ in the white matter. The new approach, therefore, led to less noisy and higher uptake enhancements than the 2D global MMA. By comparison, the use of GTM led to $+10\%$ and $-19\%$

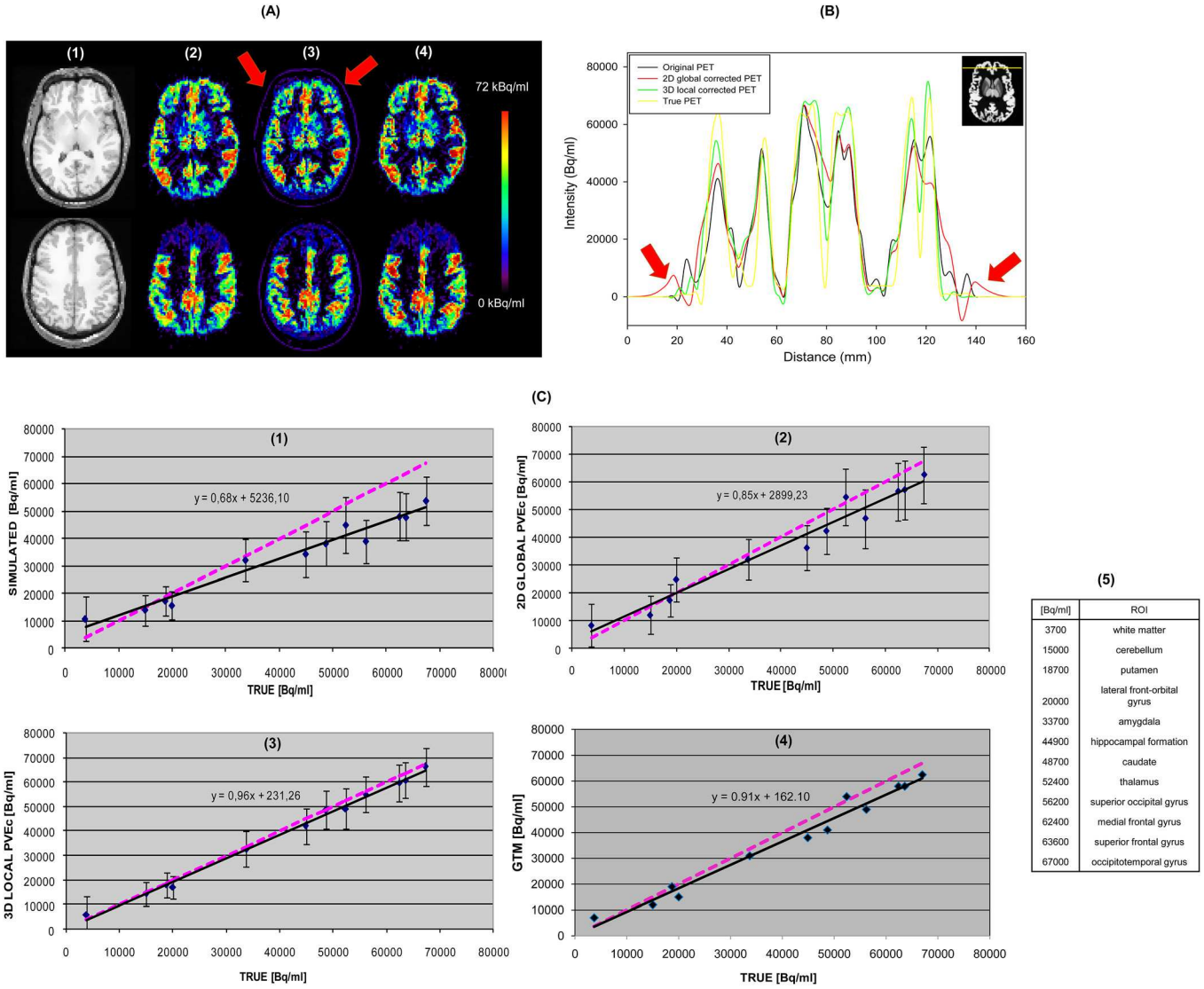



FIG. 4. (a) PET images of the simulated brain $^{18}$F-FDG PET/T1-weighted MRI dataset: (1) T1-weighted MRI, (2) noncorrected PET, (3) 2D global, and (4) 3D local MMA based PVE corrected images; (b) profile results across the frontal cortex on the uncorrected and PVE corrected PET images; (c) ROI quantification intensity (mean value ± SD) comparisons between measured (black solid lines) and true image values (magenta dotted lines): (1) for the simulated image and recovered from the corrected images using the (2) 2D global, (3) 3D local MMA, and (4) GTM approach. The ROIs and the associated true activity concentration values used in the simulation are shown in c-(5) based on the segmented T1-weighted MRI (showing highlighted a frontal region of interest).

mean activity changes for the gray and white matter ROIs, respectively.

The whole-body $^{18}$F-FDG PET/CT image [Figs. 6-(1) and 6-(2)] of a patient with lung cancer was analyzed in order to assess the potential of our new approach regarding oncology applications. Figure 6 contains the correction results using both approaches, demonstrating the incorporation of artifacts such as the spine with the global approach [Fig. 6-(3)], whereas the new model allowed avoiding them [Fig. 6-(4)]. Quantitative measurements in this bone region demonstrated an uptake increase of 28.8% ± 25.4% with the 2D global, whereas the 3D local led to a much lower variation (1.9% ± 11.3%). Table II contains the results of quantitative analysis using ROIs placed in the tumor and the lung, revealing an increase in lesion-to-lung ratio of 33.5% with the global approach, 45.8% with the GTM method and 54.1% with the 3D local approach. Furthermore, a variation in the whole lung activity concentration of 13.3% with the global approach was measured, whereas the new model leads to a smaller variation of only 6.7% thanks to the fact that it disregards anatomical structures in the lungs without matching FDG uptake. A similar variation of 5.4% of the lung uptake was obtained with the GTM approach.

## IV. DISCUSSION

Multimodality PET/CT imaging is rapidly becoming the gold standard for diagnostic studies especially in oncology with $^{18}$F-FDG PET/CT. PET/CT systems are now widely used in clinical practice thanks to the automatic fusion of functional and anatomical information they provide. Accurate and efficient PVC in this context might demonstrate

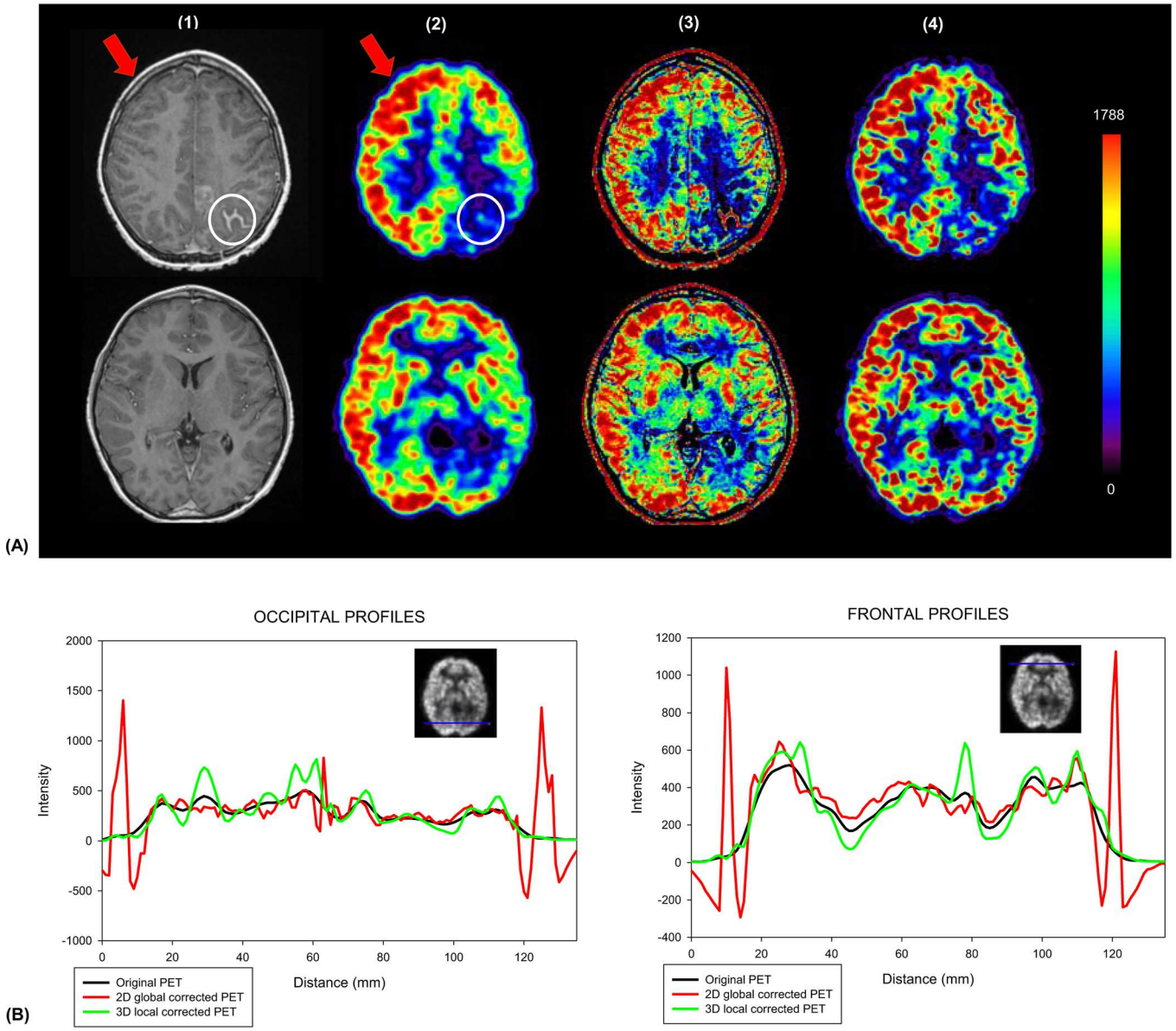



FIG. 5. (a) A clinical brain T1-weighted MRI/$^{18}$F-FDG PET with injection of gadolinium contrast: (1) T1-weighted MRI, (2) noncorrected $^{18}$F-FDG PET image, (3) 2D global, and (3) 3D local MMA based PVE corrected images; (b) profile results across the frontal and occipital cortex regions on the uncorrected and PVE corrected PET brain images corresponding to the datasets shown in (a).

significant clinical impact, for instance, for the detection of small lesions in whole-body PET acquisitions, or for the assessment of therapy response during or after treatment. In addition, new technology developments now allow the simultaneous acquisition of PET image and MRI for clinical brain studies[22] and it is expected this will be also extended to whole-body imaging. These developments facilitate the use of anatomical information either during the reconstruction (attenuation correction[23,24] or incorporation of *a priori* information[25,26]) or as a postprocessing step for the improvement of qualitative and quantitative accuracy of functional images (denoising[27] or partial volume correction[9,11]).

The recently introduced MMA methodology for partial volume correction[9] is based on the mutual multiresolution analysis of a functional image and the corresponding anatomical one. In contrast to the standard PVE correction approaches using anatomical information,[2] the MMA is voxel-wise and, hence, does not use ROIs obtained from a segmentation of the anatomical images. The algorithm was validated on synthetic and simulated datasets with accuracy similar to the reference GTM approach with the advantage of not requiring an atlas or a segmentation step as well as leading to PVE corrected images which are subsequently available for further analysis. There was, however, certain limitations associated with its 2D implementation and the use of a strictly linear and global model. This model results in the incorporation of every anatomical detail into the functional image and can therefore lead to artifacts in the corrected images where no correlation exists between anatomical and functional details. For example, the methodology has been shown to work well with a combination of FDG PET and T1 MRI brain images,[9,28] but its performance

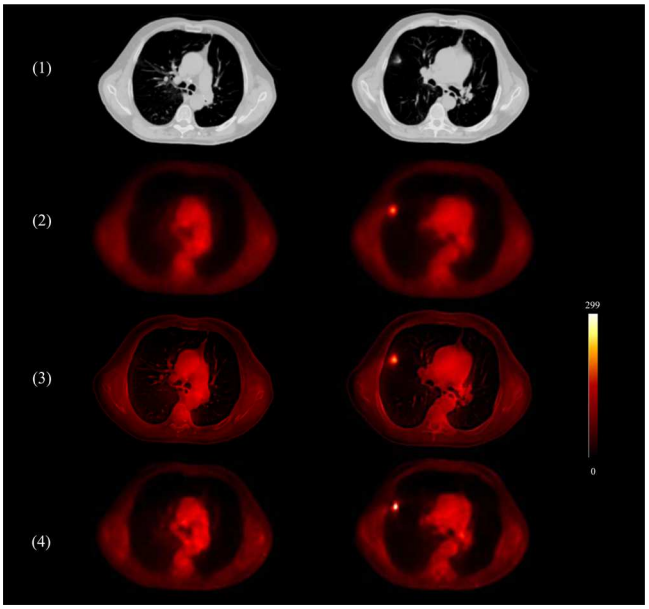



FIG. 6. A clinical whole body CT/$^{18}$F-FDG PET study: (1) CT, (2) original noncorrected PET, (3) PVE corrected images using the 2D global MMA algorithm, and (4) PVE corrected images using the 3D local MMA approach.

with respect to receptor brain PET imaging has not been demonstrated.

In this study, we have developed and evaluated a new model for MMA, based on a locally adaptive 3D analysis in order to address the limitations of the 2D global MMA previous implementation. An accurate coregistration is considered as a prerequisite to both MMA and our new proposed approach, although we also demonstrated satisfactory robustness of the method to misregistration errors. The implementation is based on the extension of the 2D analysis to the 3D and the introduction of a local model. The impact of each of the two modifications has not been evaluated individually. On the one hand, the 3D analysis provides a better representation of the PSF convolution which is a 3D phenomenon, allowing an accurate wavelet decomposition of the image compared to the 2D approach. The local analysis applied to this 3D decomposition allows computing a more accurate model which is subsequently used for the correction, compared to the simple global approach.

In addition, in order to discard uncorrelated details, the initial linear model coefficient was replaced by a new one based on the median value (rather than the mean) of the voxel-by-voxel ratios in a local cube centered on each voxel. Consequently, the 3D anatomical details in the wavelet domain are discarded in the regions where there is no matching functional uptake. The validation of the new model was performed on synthetic and simulated datasets. The results demonstrated similar quantitative performance with the 2D global approach and the GTM in the case of correlated datasets [Figs. 1(a) and 1(b)]. It is worth noting that the GTM was applied using the available ground-truth for the definition of the ROIs, leading this way to the best possible results with this approach. In addition, the performance was significantly improved for noncorrelated structural and functional image combinations [Figs. 1(c) and 4]. The new model was also evaluated using clinical whole-body and brain datasets. For both these clinical datasets we obtained a significantly improved image quality without artifacts, as well as higher contrast improvements, compared to the 2D global MMA correction (Figs. 5–6). However in the absence of ground-truth, the absolute accuracy of the recovered activity values and spatial spread of the structures of interest (such as small tumors) could not be directly assessed on these datasets.

TABLE II. ROI quantification [mean ± SD of the uptake value (kBq/ml) in different ROIs] for the original whole body PET image and the corrected one using the GTM, the 2D global and the 3D local MMA approaches. Note that there is no SD in the case of the GTM approach since it is a ROI based PVE correction approach.

| Activity (kBq/ml) | ROI lung | ROI lesion | ROI bone |
|---|---|---|---|
| Original PET | 1.5 ± 0.1 | 8.9 ± 1.4 | 5.2 ± 0.5 |
| GTM | 1.2 | 10.9 | 5.5 |
| 2D global | 1.3 ± 0.3 | 10.3 ± 1.6 | 6.7 ± 1.7 |
| 3D local | 1.4 ± 0.1 | 12.8 ± 1.9 | 5.3 ± 0.6 |

Although this new approach allows improved qualitative and quantitative voxel-wise partial volume correction using anatomical information, without assuming homogeneous uptake in regions of interest as the GTM approach, and is applicable to both brain and oncology imaging, one should consider potential pitfalls associated with any postprocessing PVE correction approach based on anatomical information. As the correction is performed based on the anatomical image details, if no or not enough information is available for a specific structure, the PVE correction will be either not possible or incomplete. The new proposed approach is certainly able to handle mismatches between anatomical and functional information. However, in the case of tumor imaging, if the lesion is necrotic in the functional image, but there is no corresponding necrosis in the anatomical structure, then the correction will be incomplete as only the external outline of the lesion will be corrected. Absence of contrast in the anatomical dataset corresponding to a specific structure would lead to a lack of significant wavelet coefficients in the wavelet decomposition of such a structure; therefore, no correction can be deduced and applied to the corresponding functional uptake. However, when a relationship exists between the anatomical signal and the functional one, such an issue may be overcome by using alternative acquisition protocols in order to generate such contrast. For instance, a contrast enhanced anatomical imaging or different sequences available in MRI imaging will be certainly interesting in enhancing the potential of PVC approaches such as our own for newly developed PET/MRI devices. Finally, one can identify the areas of the image where no significant PVE correction was applied as a result of lack of anatomical details by analyzing the MRM. This map indeed contains the correction that was applied to each voxel and very low or zero values correspond to little or no correction. It is also important to emphasize that the tissue-fraction effect can be corrected only where the frontier between tissues can be

extracted from anatomical images. This effect can in addition only be reduced, not entirely corrected, as it also exists in the anatomical datasets, although with a lower magnitude than in the functional images.

Another potential limitation of voxel-wise partial volume correction approaches based on the use of anatomical images may come from the dependency of the algorithm on the noise level in both the anatomical and functional images. However, our approach demonstrated high robustness versus noise. In addition, the local aspect of the correlation model greatly reduces the sensitivity of the correction to potential artifacts/distortions present in the anatomical images.

Misregistration is also a limiting factor for multimodality PVE correction approaches such as the one reported in this work. However, current techniques[29–31] allow fully automated 3D elastic image registration and can accurately align (with errors within 1–2 voxels) whole-body PET and CT images acquired on standalone as well as on combined PET-CT scanners. Furthermore, although both the 2D and 3D approaches are sensitive to a spatial registration error between the anatomical and functional images, the different tests carried out on synthetic images with translation movements up to 4 voxels and rotation movements up to 5°, as well as inappropriate scaling of the anatomical image showed limited impact with a maximum error of 1.9% ± 23.1% thanks to the use of the sliding cube and the median of the ratios. However, misregistration should be limited to a minimum in order to ensure the most accurate correction, since large translations and rotations might lead to errors up to 25%.

Finally, as with any postprocessing PVE correction algorithm the new model requires the exact knowledge of both modalities' spatial resolution (FHWM) and voxel sizes in order to determine the parameters of the wavelet decomposition scheme. However, our new methodology also demonstrated satisfactory robustness versus errors up to 2 mm in the choice of the FWHM parameters. After images coregistration, the process is automatic and takes from about a minute to several minutes (depending on the size of the images) on a desktop computer.

We assumed a constant value of the PSF FWHM in the entire image, which is a simplification considering the potential variation of the PSF FWHM throughout the field of view, especially in the case of whole body imaging. This aspect could be improved by modeling the exact PSF FWHM in each direction according to the spatial position of the analyzed voxel. Finally, an automatic estimation of the threshold T value regarding the wavelets coefficients might improve the results on a case by case basis. Future studies will also investigate the performance of postreconstruction processing approaches such as the one developed in this study with the incorporation of the PSF and other *a priori* information into the reconstruction iterative algorithm. Within this context a couple of existing studies have shown similar performance between the postprocessing deconvolution and PSF incorporated reconstruction based PVE correction approaches,[32,33] with the generic advantage of postreconstruction approaches being reconstruction algorithm independent.

## V. CONCLUSION

We developed an improved voxel-wise methodology to correct for partial volume effects in emission tomography. This new model overcomes limitations encountered in the originally proposed MMA and allows for a more universal approach that can potentially handle any combination of coregistered anatomical and functional images. Our new methodology extends the 2D MMA to a 3D local analysis, in which local details are conditionally taken into account in the correction process. PVE correction was successfully applied to images with either high correlation, for which the 2D MMA correction was already adequate, or more challenging cases for which correlation between anatomical and functional datasets was not complete and for which global 2D MMA failed by introducing artifacts and led to inaccurate quantification. The local 3D process was successfully tested and validated on synthetic, simulated, and clinical datasets, with similar performance to the reference GTM method without requiring a segmentation step, producing PVE corrected images, and avoiding artifacts generated by the 2D global approach. In addition, it proves to be overall more robust with good accuracy and robustness without particular requirements regarding the structural and intensity correlations between the anatomical and functional images.

[a]Author to whom correspondence should be addressed. Electronic mail: a.le-pogam@imperial.ac.uk

[1]J. A. Aston, V. J. Cunningham, M. C. Asselin, A. Hammers, A. C. Evans, and R. N. Gunn, "Positron emission tomography partial volume correction: Estimation and algorithms," J. Cereb. Blood Flow Metab. **22**, 1019–1034 (2002).

[2]O. G. Rousset, Y. Ma, and A. C. Evans, "Correction for partial volume effects in PET: Principle and validation," J. Nucl. Med. **39**, 904–911 (1998).

[3]C. C. Meltzer, P. E. Kinahan, P. J. Greer, T. E. Nichols, C. Comtat, M. N. Cantwell, M. P. Lin, and J. C. Price, "Comparative evaluation of MR-based partial-volume correction schemes for PET," J. Nucl. Med. **40**, 2053–2065 (1999).

[4]O. G. Rousset, D. L. Collins, A. Rahmim, and D. F. Wong, "Design and implementation of an automated partial volume correction in PET: Application to dopamine receptor quantification in the normal human striatum," J. Nucl. Med. **49**, 1097–1106 (2008).

[5]B. K. Teo, Y. Seo, S. L. Bacharach, J. A. Carrasquillo, S. K. Libutti, H. Shukla, B. H. Hasegawa, R. A. Hawkins, and B. L. Franc, "Partial-volume correction in PET: Validation of an iterative postreconstruction method with phantom and patient data," J. Nucl. Med. **48**, 802–810 (2007).

[6]J. Tohka and A. Reilhac, "Deconvolution-based partial volume correction in Raclopride-PET and Monte Carlo comparison to MR-based method," Neuroimage **39**, 1570–1584 (2008).

[7]N. Boussion, C. Cheze Le Rest, M. Hatt, and D. Visvikis, "Incorporation of wavelet-based denoising in iterative deconvolution for partial volume correction in whole-body PET imaging," Eur. J. Nucl. Med. Mol. Imaging 36, 1064–1075 (2009).

[8]A. S. Kirov, J. Z. Piao, and C. R. Schmidtlein, "Partial volume effect correction in PET using regularized iterative deconvolution with variance control based on local topology," Phys. Med. Biol. **53**, 2577–2591 (2008).

[9]N. Boussion, M. Hatt, F. Lamare, Y. Bizais, A. Turzo, C. Cheze-Le Rest, and D. Visvikis, "A multiresolution image based approach for correction of partial volume effects in emission tomography," Phys. Med. Biol. **51**, 1857–1876 (2006).

[10]P. Dutilleux, "An implementation of the 'algorithme a trous' to compute the wavelet transform," in *Time-Frequency Methods and Phase Space*, edited by Springer-Verlag (Springer-Verlag, Marseille, France, 1987), Vol. 1, pp. 298–304.

[11]M. Shidahara, C. Tsoumpas, A. Hammers, N. Boussion, D. Visvikis, T. Suhara, I. Kanno, and F. E. Turkheimer, "Functional and structural synergy for resolution recovery and partial volume correction in brain PET," Neuroimage **44**, 340–348 (2009).
[12]I. Daubechies, "The wavelet transform, time-frequency localization and signal analysis," IEEE Trans. Inf. Theory **36**, 961–1005 (1990).
[13]S. Mallat, "A theory for multiresolution signal decomposition: The wavelet representation," IEEE Trans. Pattern Anal. Mach. Intell. **11**, 674–693 (1989).
[14]D. A. Yocky, "Artifacts in wavelet image merging," Opt. Eng. **53**, 2094–2101 (1995).
[15]Krista Amolins *et al*., "Wavelet based image fusion techniques—An introduction, review and comparison," ISPRS J. Photogramm. Remote Sens. **62**, 249–263 (2007).
[16]M. J. Shensa, "The discrete wavelet transform: Wedding the a trous and Mallat algorithms," IEEE Trans. Inf. Theory **40**, 2464–2482 (1992).
[17]J. L. Starck, J. Fadili, and F. Murtagh, "The undecimated wavelet decomposition and its reconstruction," IEEE Trans. Image Process. **16**, 297–309 (2007).
[18]M. Hatt, C. Cheze le Rest, A. Turzo, C. Roux, and D. Visvikis, "A fuzzy locally adaptive Bayesian segmentation approach for volume determination in PET," IEEE Trans. Med. Imaging **28**, 881–893 (2009).
[19]I. G. Zubal, C. R. Harrell, E. O. Smith, Z. Rattner, G. Gindi, and P. B. Hoffer, "Computerized three-dimensional segmented human anatomy," Med. Phys. **21**, 299–302 (1994).
[20]C. Tsoumpas, F. E. Turkheimer, and K. Thielemans, "Study of direct and indirect parametric estimation methods of linear models in dynamic positron emission tomography," Med. Phys. **35**, 1299–1309 (2008).
[21]J. Ashburner and K. J. Friston, "Voxel-based morphometry—The methods," Neuroimage **11**, 805–821 (2000).
[22]H. P. Schlemmer, B. J. Pichler, M. Schmand, Z. Burbar, C. Michel, R. Ladebeck, K. Jattke, D. Townsend, C. Nahmias, P. K. Jacob, W. D. Heiss, and C. D. Claussen, "Simultaneous MR/PET imaging of the human brain: Feasibility study," Radiology **248**, 1028–1035 (2008).
[23]P. E. Kinahan, D. W. Townsend, T. Beyer, and D. Sashin, "Attenuation correction for a combined 3D PET/CT scanner," Med. Phys. **24**, 2046–2053 (1998).
[24]D. Visvikis, D. C. Costa, I. Croasdale, A. H. Lonn, J. Bomanji, S. Gacinovic, and P. J. Ell, "CT-based attenuation correction in the calculation of semi-quantitative indices of [18F]FDG uptake in PET," Eur. J. Nucl. Med. Mol. Imaging **30**, 344–353 (2003).
[25]C. Comtat, P. E. Kinahan, J. A. Fessler, T. Beyer, D. W. Townsend, M. Defrise, and C. Michel, "Clinically feasible reconstruction of 3D whole-body PET/CT data using blurred anatomical labels," Phys. Med. Biol. **47**, 1–20 (2002).
[26]J. Nuyts, K. Baete, D. Beque, and P. Dupont, "Comparison between MAP and postprocessed ML for image reconstruction in emission tomography when anatomical knowledge is available," IEEE Trans. Med. Imaging **24**, 667–675 (2005).
[27]F. E. Turkheimer, N. Boussion, A. N. Anderson, N. Pavese, P. Piccini, and D. Visvikis, "PET image denoising using a synergistic multiresolution analysis of structural (MRI/CT) and functional datasets," J. Nucl. Med. **49**, 657–666 (2008).
[28]N. Boussion, C. Cheze-Le Rest, Y. Bizais, and D. Visvikis, "Quantitative assessment by means of realistic simulated images and patient data of a new method for partial volume correction effects in brain PET," J. Nucl. Med. **47**, 192P (2006).
[29]R. Shekhar, V. Walimbe, S. Raja, V. Zagrodsky, M. Kanvinde, G. Wu, and B. Bybel, "Automated 3-dimensional elastic registration of whole-body PET and CT from separate or combined scanners," J. Nucl. Med. **46**, 1488–1496 (2005).
[30]C. O. Sorzano, P. Thevenaz, and M. Unser, "Elastic registration of biological images using vector-spline regularization," IEEE Trans. Biomed. Eng. **52**, 652–663 (2005).
[31]Frost and Sullivan, *Emerging Technology Developments in Fusion Technology for Diagnostic Imaging* (Technical Insights, San Antonio, Texas, 2005).
[32]A. Le Pogam, A. M. Alessio, B. Kandel, A. S. Kirov, O. Barrett, P. Fernandez, C. Cheze Le Rest, and D. Visvikis "Comparison of PET reconstruction resolution recovery and post-reconstruction deconvolution for PET partial volume correction," J. Nucl. Med. **51**(Suppl 2), 577 (2010)
[33]N. J. Hoetjes, F. H. van Velden, O. S. Hoekstra, C. J. Hoestra, N. C. Krak, A. A. Lammertsma, and R. Boellaard, "Partial volume correction strategies for quantitative FDG PET in oncology," Eur. J. Nucl. Med. Mol. Imaging **37**, 1679–1687 (2010).

Multi-observation PET image analysis for patient follow-up quantitation and therapy assessment



2011 Phys. Med. Biol. 56 5771

(http://iopscience.iop.org/0031-9155/56/18/001)

IOP PUBLISHING PHYSICS IN MEDICINE AND BIOLOGY

Phys. Med. Biol. **56** (2011) 5771–5788 doi:10.1088/0031-9155/56/18/001

# Multi-observation PET image analysis for patient follow-up quantitation and therapy assessment

**S David[1,3], D Visvikis[1], C Roux[1,2] and M Hatt[1]**

[1] INSERM U650, LaTIM, Brest, F-29200, France
[2] Institut Telecom-Telecom Bretagne, Brest, F-29200, France

E-mail: simon.david@etudiant.univ-brest.fr

Received 10 March 2011, in final form 1 July 2011
Published 16 August 2011
Online at stacks.iop.org/PMB/56/5771

**Abstract**

In positron emission tomography (PET) imaging, an early therapeutic response is usually characterized by variations of semi-quantitative parameters restricted to maximum SUV measured in PET scans during the treatment. Such measurements do not reflect overall tumor volume and radiotracer uptake variations. The proposed approach is based on multi-observation image analysis for merging several PET acquisitions to assess tumor metabolic volume and uptake variations. The fusion algorithm is based on iterative estimation using a stochastic expectation maximization (SEM) algorithm. The proposed method was applied to simulated and clinical follow-up PET images. We compared the multi-observation fusion performance to threshold-based methods, proposed for the assessment of the therapeutic response based on functional volumes. On simulated datasets the adaptive threshold applied independently on both images led to higher errors than the ASEM fusion and on clinical datasets it failed to provide coherent measurements for four patients out of seven due to aberrant delineations. The ASEM method demonstrated improved and more robust estimation of the evaluation leading to more pertinent measurements. Future work will consist in extending the methodology and applying it to clinical multi-tracer datasets in order to evaluate its potential impact on the biological tumor volume definition for radiotherapy applications.

(Some figures in this article are in colour only in the electronic version)

## 1. Introduction

Positron emission tomography (PET) is now a widely used tool in the field of oncology, especially in applications such as diagnosis, patient follow-up studies (Krak *et al* 2005)

[3] Author to whom any correspondence should be addressed.

or in radiotherapy planning (Jarritt *et al* 2006). In the context of patient follow-up, early metabolic changes detected with 2′-deoxy-2′-[$^{18}$F]-fluoro-D-glucose (FDG) PET imaging can occur before anatomic changes observed with computed tomography (CT) imaging. By assessing differences in several PET scans acquired before and at different times during treatment, various qualitative and quantitative methods have been proposed to characterize the therapeutic response (Weber 2007). In patient monitoring studies, qualitative methods such as visual assessment are less accurate and reproducible than quantitative measurements (Lin *et al* 2007). Furthermore, different therapeutic parameters (indexes) have been defined either on dynamic or static PET acquisitions with a similar reproducibility (Weber *et al* 1999). Being less restrictive in clinical routine, only the parameters computed in the static PET scans have been considered in our work. Most widely used in patient follow-up studies, the standardized uptake value (SUV) measures the tracer uptake in the tumor. Derived from the SUV index, two measurements, namely the maximum SUV ($SUV_{max}$) and the mean SUV ($SUV_{mean}$), were assessed in our study by computing, respectively, the maximum and the mean of SUV in voxels included in a region of interest defining the tumor. The reproducibility and the robustness of both SUV indexes have been previously assessed (Weber 2007, Nahmias and Wahl 2008) and compared to the reproducibility of tumor volume measurements with various automated methodologies (Hatt *et al* 2010). An early therapeutic response can be characterized by measuring relative or absolute SUV variations between pre-treatment and mid-treatment PET scans. Other quantitative parameters have been used such as the total lesion glycolysis (TLG) defined as the product of the mean SUV and the tumor volume (Larson *et al* 1999, Hatt *et al* 2010).

The therapeutic response is usually estimated by measuring the tumor size on the CT scans, and according to guidelines such as World Health Organization (1979) and RECIST (Therasse *et al* 2000). More recent criteria have been proposed such as PERCIST (Wahl *et al* 2009), adding the consideration of quantitative parameters extracted from PET images. However, these criteria are still limited to simple SUV measurements and do not include volumetric characterization of the tumors, and no guidelines have been established recommending the best way to characterize the therapy response according to the variation of metabolically active tumor volumes. In the current clinical practice, the therapeutic response is therefore usually assessed by considering one single value as the $SUV_{max}$ within the primary lesion, extracted from each PET scan. Presently, the measure of $SUV_{max}$ variation is considered as the gold standard of the treatment response definition. This method however accounts neither for the tumor volume variations nor the spatial uptake variation within the tumor volume.

Among the new methodologies developed in PET tumor delineation (Zaidi and El Naqa 2010), most of them have only considered the use of such delineation for static images segmentation and for diagnosis/prognosis. A few authors have recently proposed different methodologies dedicated to PET follow-up, like the one by Necib *et al* (2008), which is aimed at assessing a response by comparing two follow-up PET images. After voxel-to-voxel registration of the two scans, a biparametric map is generated representing the tracer uptake variations within the tumor. In the context of cancer treatment prediction, El Naqa *et al* (2009) have recently proposed a texture-based approach, considering texture properties of voxels within tumors as prognosis factors for the assessment of therapy response.

Regarding the use of PET in radiotherapy, the gross tumor volume (GTV) definition is usually carried out manually on fused FDG-PET/CT scans. However, imaging tumor's glucose consumption with the FDG alone may not be sufficient to determine the GTV (Mankoff *et al* 2003). Considering the measure of other features of cancer metabolism like proliferation, hypoxia and apoptosis using additional tracers may generate more complete information regarding the target tumor volume (Bentzen 2005, Shields 2003, Vaupel and

Mayer 2007). Accurate tumor volume delineation would therefore require a fusion of all available measurements obtained with these different tracers. Such a fusion could be valuable to thoroughly and potentially more accurately assess tumor volume definition as well as evolution during therapy.

The main objective of this study was to develop a fusion method derived from multi-observation approaches such as these developed in satellite and astronomical imaging (Masson and Pieczynski 1993). Considering either patient follow-up and/or multi-tracer PET datasets, the proposed method aims at assessing a treatment response and tumor volume definition by automatically determining the different variations of tracer uptakes within the regions of the analyzed fused images. Our approach is statistical and assumes that the data can be modeled by a mixture distribution of multi-observation random fields. The parameters defining the mixture distribution are estimated using a stochastic expectation maximization (SEM) algorithm (Celeux and Diebolt 1986) combined with a locally adaptive spatial priors estimation in order to account for voxels correlation.

Our method was applied to simulated and clinical pre- and post-treatment PET scans of esophageal cancer within the context of radio-chemotherapy follow-up. It was compared to current quantitative methods proposed for the assessment of the therapeutic response based on tumor volume evolution, namely the definition of the tumor volumes independently on both scans using adaptive thresholding.

## 2. Materials and methods

### *2.1. Multi-observation framework*

The proposed method could potentially be applied for both patient follow-up applications and multi-tracer analysis using PET scans. The proposed method is aimed at merging the available PET images in order to derive a fusion of the information regarding the treatment response in patient follow-up application or/and the multi-tracer tumor volume, as illustrated in figure 1. While the analysis for both applications might require different fusion rules or interpretation, the basics of the approach are the same and are based on the unsupervised Bayesian methods, widely used in segmentation and classification of satellite, astronomical or medical imaging (Masson and Pieczynski 1993, Pieczynski 2003, Hatt *et al* 2009).

*2.1.1. Bayesian model.* Let $T$ be a finite set corresponding to the voxels of 3D registered PET images. We consider two random processes $\mathbf{Y} = (\boldsymbol{y_t})_{t \leqslant T}$ and $\mathbf{X} = (x_t)_{t \leqslant T}$. $\mathbf{Y}$ models the observed multi-tracers or follow-up PET scans, acquired at different times during the treatment, and takes its values in $\mathbb{R}$. Each $\boldsymbol{y_t}$ is therefore a vector of real values defined as $\boldsymbol{y_t} = \left(y_t^{(1)}, \ldots, y_t^{(B)}\right)$, containing the voxel values of each PET image, with $B$ being the image number observed in the fusion. Each $\boldsymbol{y_t}$ is associated with a label $x_t$. $\mathbf{X}$ models the fusion map which is designated in our specific application as the therapeutic response classification. $\mathbf{X}$ takes its values in a set $\{1 \ldots K\}$, with $K$ being the number of classes that is usually user-dependent and defined depending on the fusion goal. The objective of the approach is therefore to estimate the distribution of $(\mathbf{X}, \mathbf{Y})$. Considering the Bayesian framework, the relationship between $\mathbf{X}$ and $\mathbf{Y}$ can be modeled using the joint probability:

$$p(\mathbf{X}, \mathbf{Y}) = p(\mathbf{Y}|\mathbf{X}) \times p(\mathbf{X}), \tag{1}$$

where $p(\mathbf{X})$ is the prior knowledge about $\mathbf{X}$ and $p(\mathbf{Y}|\mathbf{X})$ is the 'noise model': the likelihood of the observation $\mathbf{Y}$ conditionally to the hidden ground-truth $\mathbf{X}$. In this Bayesian framework, the prior knowledge $p(\mathbf{X})$ can be modeled globally, for instance by considering Markovian models

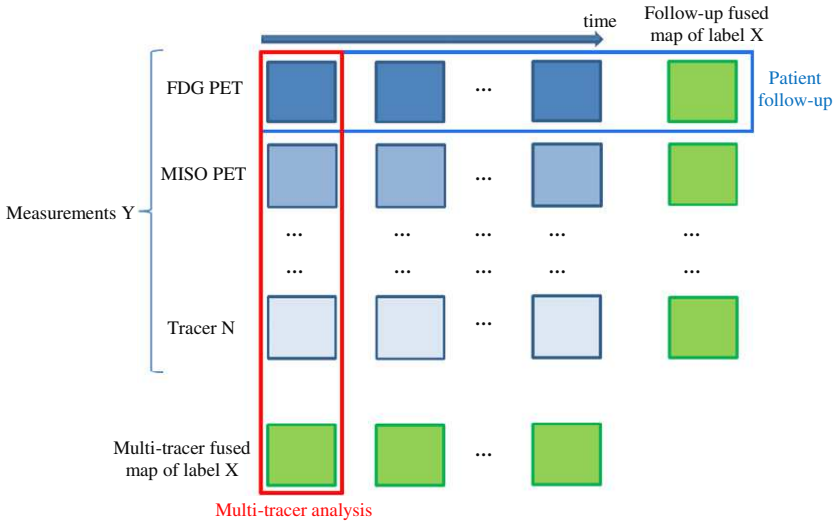


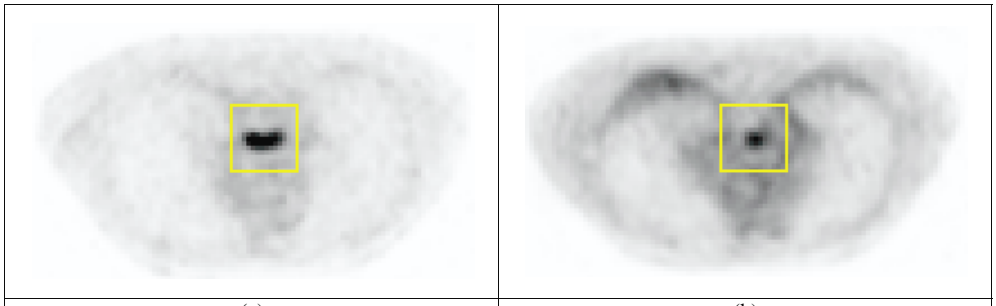

**Figure 1.** Multi-observation framework of multi-tracer and patient follow-up data.

(Pieczynski 2003) such as chains or trees, or locally with blind, contextual or adaptive models (Peng and Pieczynski 1995). Different noise distributions can be used for the observation model $p(\mathbf{Y}|\mathbf{X})$ such as Gaussian or generalized Gaussian (Delignon *et al* 1997).

In our fusion method, we have used a locally adaptive prior model and the noise model has been assumed Gaussian; however, other distributions could be considered in future developments. The distribution of $(\mathbf{X}, \mathbf{Y})$ is hence defined by the priors $\Pi_k$, the mean vectors $\boldsymbol{\mu_k}$ and covariance matrices $\Gamma_k$ associated to each of the $K$ classes in the mixture. The mean parameter $\boldsymbol{\mu_k}$ is a vector associated to the $B$ images of the fusion $\boldsymbol{\mu_k} = \left(\mu_k^{(1)}, \ldots, \mu_k^{(B)}\right)$. The SEM algorithm was used here to estimate the parameters of the distribution of $(\mathbf{X}, \mathbf{Y})$. It is a stochastic version of the classic EM algorithm, ensuring better and faster convergence as well as higher independence on the initialization. From here onwards, our approach will be denoted as ASEM.

### 2.2. *Fusion process*

*2.2.1. Pre-processing: image deconvolution.* PET images are characterized by their high level of noise and the limited spatial resolution inducing partial volume effect (PVE) (Soret *et al* 2007). The under estimation of the tissues uptakes and activity cross contamination between structures with different uptakes are two consequences of the PVE effects in PET images. When considering several co-registered PET images, the voxels most affected by PVE may not be on the same coordinates for each scan, which implies intensities distributions that might complexify both estimation and classification steps in the fusion process. In addition, without PVE correction (PVC), SUV values extracted from each scan may be significantly biased. This can lead to under or over estimation of the uptake variation between pre- and post-treatment scans, especially if significant tumor volume variation occurs. Indeed, PVE impact on the SUV measurement within the tumor strongly depends on the object's size. In order to reduce the impact of these effects on the subsequent steps, a PVE correction (PVC) was applied to each image prior to their fusion. The chosen PVC method was developed by Boussion *et al* (2008) and further improved by Le Pogam *et al* (2009) and consists of a 3D



(a) (b)

**Figure 2.** (a) Illustration of VOI definition in the pre-treatment scan and (b) automatically reported on the registered mid-treatment scan.

voxel-wise correction using an iterative deconvolution improved by wavelet-based optimal denoizing of the residual. This preprocessing step offers two advantages: first, it reduces the size of blurred frontiers between the different regions of the images, hence reducing their impact on subsequent registration and fusion complexity. Second, it allows the extraction of corrected uptake values from fusion maps for a quantitative characterization of evolution of the activity within tumor volume and/or sub-volumes.

*2.2.2. Local-based analysis.* As our goal is to automatically determine the variation of tracer activity and position/volume of a functional tumor, we assume that the overall tumor volume has been previously automatically or manually isolated in a 3D box or volume of interest (VOI) by a clinician on the co-registered PET images. Therefore, the box should be large enough to encompass the entire tumor in each scan and avoid including too many neighboring tissues with significant physiological uptake. Consequently, the definition of such a processing box should allow any shape and size in 3D. The definition of this 3D VOI should therefore be carried out on the scan in which the tumor appears to be the largest, and automatically registered on the other volumes involved in the fusion as illustrated in figure 2.

*2.2.3. Fuzzy k-means initialization and choice of the number of classes.* In unsupervised Bayesian segmentation framework, the initialization is an important step. In our method, we used the fuzzy $k$-means (FKM) algorithm (Krishnapuram and Keller 1994) based on fuzzy logic applied to the voxels values. In his PhD thesis, Provost (2001) described an improved version of FKM, allowing the automatic estimation of the optimal number of classes in the mixture, based on the use of an entropy criterion and a user selection of the upper limit of the number of classes. This upper limit was defined as the product of the number of images and the number of classes within each image considered in the fusion. In each iteration and for all the voxels of the image, a membership coefficient associated to the $K$ classes of the mixture is estimated. At the end of FKM execution, the mixture parameters associated with each class $k$ ($\Pi_k$, $\boldsymbol{\mu_k}$, $\Gamma_k$) are initialized for the Bayesian estimation. The cost function of the FKM algorithm is defined as

$$E = \sum_{k=1}^{K} \sum_{t=1}^{T} (\psi_{kt})^d \, \|x_t - c_k\|^2 + \alpha \sum_{k=1}^{K} p_k^2 \tag{2}$$

with

$$\sum_{k=1}^{K} \psi_{kt} = 1, \quad \forall t \quad \text{and} \quad p_k = \sum_{t=1}^{T} \psi_{kt}, \tag{3}$$

where $E$ is the cost function to minimize, $K$ is the number of classes in the mixture, $T$ is the number of voxels in the volume, $\psi_{kt}$ is the membership matrix of each voxel $t$ to the $k$th class, $c_k$ is the centroid of the $k$th class. The second term of the equation (2) is the entropic criterion where $\alpha$ is the parameter weighting the cost function. The minimization of the FKM cost function is performed by constraining the entropic term. This latter is high in the first iterations, leading to a reduction of the number of classes in the mixture, and it decreases exponentially in order to allow the FKM classification process.

*2.2.4. Parameters estimation.* The parameters $(\Pi_k, \boldsymbol{\mu_k}, \Gamma_k)$ defining the Gaussian mixture of the $(\mathbf{X}, \mathbf{Y})$ distribution are estimated by the SEM algorithm by sampling several realizations of $\mathbf{X}$ according to its posterior distribution $p(\mathbf{X}|\mathbf{Y})$. In the adaptive framework, the global prior $\Pi_k$ associated to the $k$th class of the mixture are re-estimated using a local neighboring 3D cube and replaced by local priors $\pi_{t,k}$ defined for each voxel and each class. The mean vector $\boldsymbol{\mu_k}$ and the covariance matrix $\Gamma_k$ are finally computed for each of the $K$ classes in the mixture. The details of parameters estimation with the SEM algorithm are given in the appendix section.

*2.2.5. Decision step.* In order to perform fusion on a voxel-by-voxel basis, we used a classification criterion to assign a class to each voxel. For this purpose we chose the *maximum likelihood* method. To compute a solution, this criterion requires the parameters defining the *a priori* model (priors of each class and for each voxel) as well as the observation data model (mean and covariance matrices of each class), previously estimated using the SEM algorithm (see appendix).

### *2.3. Simulated datasets*

In order to evaluate the behavior of the fusion approach within the context of tumor evolution assessment, we considered realistic simulations of non-spherical tumors. These simulated tumors were created using as models real head and neck and esophageal observed in clinical datasets. The procedure for the simulations of such data has been previously described (Le Maitre *et al* 2009). The simulated cases used in our study are composed of two simulated PET scans, one before and one after the treatment. The clinical cases used as models corresponded to patients classified as partial responders or progressive disease to the radiochemotherapy according to RECIST criteria.

In order to evaluate the robustness of the methods, three levels of noise were considered for every simulated acquisition by selecting 100, 80 and 60% of the simulated lines of response for the iterative reconstruction, respectively. With 20 clinical follow-up cases and considering three levels of noise and various tumor-to-background ratios for each case, 70 different simulated cases were generated. Most of the simulated datasets, representing 15 out of the 20 cases, were generated from patients classified as partial responders. The others cases were designed to simulate progressive disease. The mean tumor volume and tumor to background ratio used in the first and second simulated follow-up cases are given in table 1. Three examples of these simulated follow-up cases (showing only central axial slice) with their associated ground-truth are illustrated in figure 3. Similar to the clinical datasets they are based on, these tumors are characterized by either homogenous or heterogeneous tracer uptake. Background activity was simulated as homogeneous. The voxels were assumed to

**Table 1.** Mean tumor to background ratios and mean tumor volumes computed for all the simulated follow-up cases.

| | Tumor volume ($cm^3$) | | T/B ratio | |
|---|---|---|---|---|
| PET 1 | $34.1 \pm 27$ | (6–90) | $5.7 \pm 2$ | (2.7–9.3) |
| PET 2 | $17.4 \pm 2.5$ | (1.9–101) | $4.2 \pm 1.4$ | (2.0–6.5) |

belong either to the background (BD) or to the tumor (T). In these simulated images, no registration was required.

### *2.4. Clinical datasets*

*2.4.1. Patient data.* After preliminary studies on simulated follow-up PET scans, the fusion method was tested on real clinical datasets. Seven patients with esophageal cancer undergoing concomitant radiochemotherapy between 2005 and 2008 were considered, with one PET scan before treatment, and another PET scan after treatment, both acquisitions carried out according to the same protocol. All these patients were classified as partial responders one month after the completion of treatment, according to RECIST. Consequently, the variation of metabolically active tumor volume as well as the SUV within the tumor volume is expected to be less than 100% since residual tumor uptake is seen for all these patients, and also above 20 to 30% which is their reproducibility limits as previously demonstrated (Hatt *et al* 2010). No volume or uptake increase should be measured in these cases. Visual illustrations of three clinical follow-up cases are given in figure 5. The physiological uptake of the mediastinum around the tumor volume was much more significant in the clinical images than in simulated cased (for which background was simulated as homogeneous) and was therefore taken into account in the fusion process as an additional class. Thus, on each scan, the voxels were assumed to belong to the background (BD) of the lungs, the physiological tissues (PHY) of the mediastinum or the tumor volume (T).

*2.4.2. PET/PET registration.* In the context of patient follow-up, the PET/CT images are acquired at several month intervals. As an important prerequisite of the proposed method, which works on a voxel level, the scans must therefore be registered before the fusion method can be applied. The PET/PET registration was carried out using a method previously proposed (Ouksili *et al* 2007), in which the PET data are first registered with their associated CT scans, acquired in the same bed position. Having more landmarks and a higher resolution, the CT scans are registered using the MIPAV software. The CT/CT registration was carried out using a rigid transformation, which optimizes the least-squares criterion of a large VOI. The CT/CT transformation matrix was then applied to the PET scans for registration. A rigid registration was considered to be sufficient since the procedure was carried out only on small 3D regions of interest surrounding the lesions which were located in the mediastinum or head and neck regions. In addition, a rigid transformation avoids the deformation of tumor volumes in the PET images which would be certainly associated with the use of a deformable model. Finally, the consideration of head and neck and mediastinum lesions reduces the potential influence of respiratory motion on the registration process.

In the fusion process, the use of a sliding estimation cube as described in section 2.2 for the computation of the *spatial priors* is expected to reduce the impact of small registration errors of the PET datasets. The impact of the scans misregistration on the fusion process was considered in this study by shifting the second scan in a subgroup of the simulated datasets

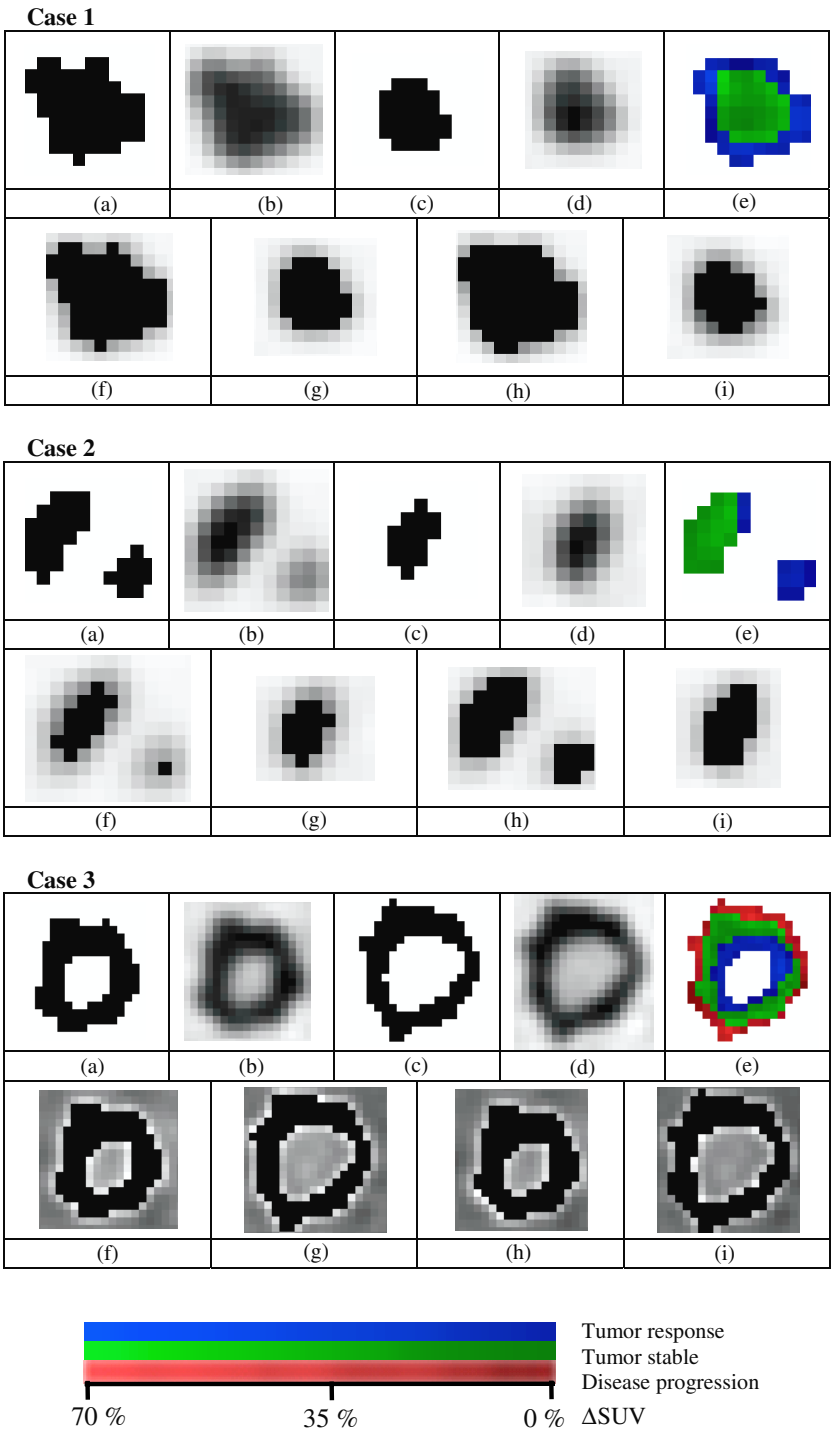



**Figure 3.** (b) and (d) Simulated follow-up tumors, (a) and (c) associated ground-truths, (e) ASEM fusion map, individual segmented map with the adaptive threshold (f) and (g), and the ASEM method (h) and (i) for the three simulated cases.

by one and two voxels (4 and 8 mm, respectively) in a random direction, and then quantifying the impact on the resulting volume error for each scan after ASEM fusion. The one to two voxel shifts used in this work corresponds to typical registration errors associated with the use of fusion algorithms in CT imaging.

*2.4.3. Quantitative measurement of clinical datasets.* In patient monitoring studies, the SUV is the most used semi-quantitative parameter and is defined as

$$\mathrm{SUV} = \frac{C}{A \times W}, \tag{4}$$

where $C$ is the tracer concentration, $A$ the injected activity and $W$ the patient weight. Among the different SUV indexes available, the most often used in clinical practice are $\mathrm{SUV}_{\mathrm{max}}$ and $\mathrm{SUV}_{\mathrm{mean}}$, computed, respectively, as the highest value and the mean of voxels values in a given region of interest (ROI), usually defined manually. Contrary to the $\mathrm{SUV}_{\mathrm{max}}$, the computation of the $\mathrm{SUV}_{\mathrm{mean}}$ depends on the volume of the ROI. Our fusion method allows identifying the variations of concentration activities. Therefore, according to the ASEM fused map, an estimation of each individual tumor volume can be carried out. Then, for each scan, a measure of $\mathrm{SUV}_{\mathrm{mean}}$ can be extracted according to the metabolic volume of the tumor $V_{\mathrm{ASEM}}$ estimated with the multi-observation method. Note that the SUV values extracted from the fusion maps are values corrected for PVE due to the deconvolution pre-processing step.

### *2.5. Alternative approaches used for comparison*

We compared the multi-observation fusion results with methods that have been proposed for patient follow-up studies. Most clinical studies only consider $\mathrm{SUV}_{\mathrm{max}}$ variation. In order to take into account full metabolically active volumes evolution as a response criterion, it has been suggested to determine them independently on both scans using threshold-based methodologies. Many studies have demonstrated that a fixed threshold value not adequate for this task and that adaptive thresholding taking into account the background uptake performs better (Nestle *et al* 2005, Tylski *et al* 2010). Manual tumor delineation by experts was not considered in the study, due to its high inter- and intra-observer variability (Hatt *et al* 2010). We compared our approach to independent delineation of tumors in each scan using adaptive threshold, the value of which is determined from the estimated contrast between the tumor activity and the background activity and is optimized for a given scanner using phantom acquisitions of spheres (Erdi *et al* 1997). Such an optimization was performed for our acquisition protocol and scanner model. An associated $\mathrm{SUV}_{\mathrm{mean}}$ value was also computed.

### *2.6. Evaluation metric for simulated datasets*

True volumes of simulated tumors are known. Therefore, the assessment of the fusion process was achieved by the estimation of volume errors (VE). For each simulated case, segmented maps of the first and second follow-up scans are obtained with the two methods, as illustrated in figure 3. The individual segmented maps of the multi-observation method can actually be deduced from the ASEM fused map. Although volume errors may be larger than 100% in specific cases for which delineation completely fails, errors were limited to 100%.

### *2.7. Quantitative variation for clinical datasets*

No ground-truth was available for the clinical follow-up cases. Hence, to compare the methods, the following quantitative indexes have been considered. The variation of metabolic tumor

volume ($\Delta V$) and mean of SUV ($\Delta SUV_{mean}$) between the pre-treatment and the follow-up PET scans for the different methods were measured with the two methods. In addition, the evolution of the original (without PVE correction) maximum SUV in the ROI was considered for comparison as it is the one currently used in clinical practice in oncology and defined as the gold standard. Contrary to the simulated datasets, the background level in clinical cases is not homogenous due to the physiological uptake of the mediastinum. The intra-observer variability of the adaptive thresholding method has been investigated for the same context of esophageal lesions in a previous study and demonstrated significant variability (Hatt *et al* 2011). Therefore, the adaptive threshold segmentation was carried out here by two clinicians with similar training and experience for each follow-up case, in order to evaluate the impact on the measurement of quantitative variations. The two clinicians followed a specific protocol: they were instructed to measure the mean background value by placing manually a ROI within the mediastinum, at least a few cm away from the lesions. They were free to choose the exact location and size of the ROI.

## 3. Results

### *3.1. Results on simulated datasets*

For the selected simulated cases, the fusion maps obtained by our multi-observation method are illustrated in figure 3(e). In order to facilitate the interpretation of the ASEM fused maps, colors have been affected to the different uptake variations. Blue areas represent a response (negative difference in tracer uptake between the two scans) whereas green color is associated with a stable tumor (similar significant uptake in both scans at this location). Red color was used to indicate tumor progression (higher uptake in second scan with respect to first one). Note that ASEM never wrongly resulted in tumor progression or regression. The intensity associated with each voxel in the fusion map is set as the SUV relative variation ($\Delta SUV$) between the first and the second scan. The segmented maps of the first and second chosen follow-up scan computed the different methods are presented in figure 3(f–h).

Mean volume errors and associated standard deviation associated with the use of ASEM fusion or independent adaptive threshold-based delineation, for all pre-treatment and post-treatment simulated cases are presented in figure 4. The VE computed for the first follow-up scan was significantly (Kruskal-Wallis tests $p < 0.0001$) lower for ASEM ($-2.6 \pm 8\%$) than for the adaptive threshold ($+28 \pm 17\%$). For the second follow-up scan, however, the VE were higher for both methods. Adaptive threshold led to higher overestimation of the tumor volumes than for the pre-treatment image ($+30 \pm 15\%$), whereas ASEM led to underestimation of the true post-treatment tumor volume ($-9 \pm 25\%$) with a larger variability. Adaptive threshold gave large volume overevaluation in all cases, whereas ASEM led to better results in several cases but higher errors in some other cases. In both pre- and post-treatment images, the adaptive threshold method tended to overestimate the tumor volume, with larger absolute errors than the ASEM method, that tended to underestimate the volume in the second scan.

Three selected cases among the simulated dataset are illustrated in the figure 3. For the first case, no significant differences were observed between the use of the adaptive threshold applied independently to each image and the ASEM method. For the second more complex case, both methods showed different results. As illustrated in figure 3(f), the adaptive threshold led to an underestimation of the tumors uptake, contrary to the ASEM method. As the fusion map (figure 3(e)) shows, the disappearing lesion is correctly identified (in blue) whereas the larger lesion is shown as stable (in green) with a small blue part, indicating that this

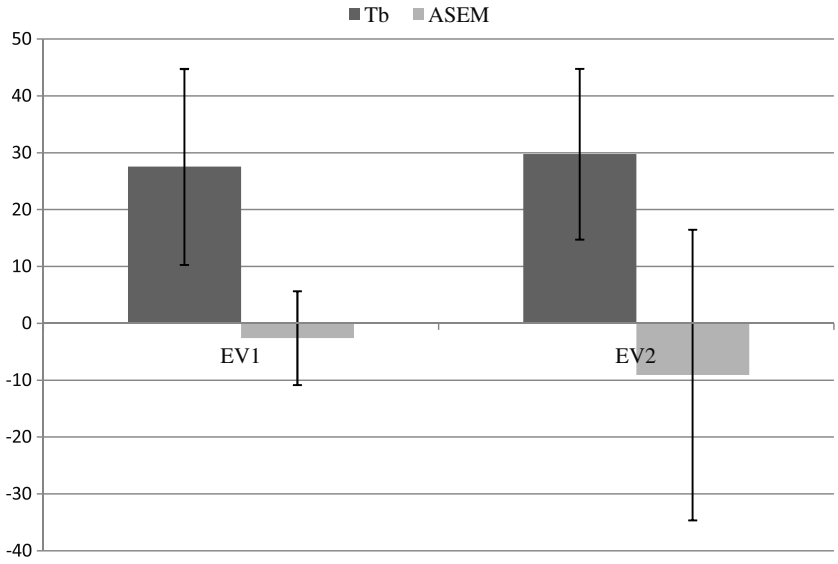



**Figure 4.** Mean VE (%) with standard deviation as error bars of the first and second follow-up scans for adaptive threshold and ASEM methods applied to the simulated cases.

tumor uptake has indeed smaller volume in the post-treatment scan. The third case illustrates the evolution of a necrotic tumor, on which ASEM correctly identified the various parts of evolution (blue, green, red).

Finally, the ASEM fusion proved robust to random misregistration of one voxel (4 mm) with volume errors increasing from $-2.6 \pm 8\%$ to $-9 \pm 16\%$ and $-9 \pm 25\%$ to $-13.5 \pm 30\%$ for first and second scan, respectively. A spatial shift of two voxels (8 mm) led to similar error levels regarding the first scan ($-9 \pm 17\%$ versus $-9 \pm 16\%$) but a higher increase regarding the second scan ($-21 \pm 37\%$ versus $-13.5 \pm 30\%$).

### *3.2. Results on clinical datasets*

The fusion maps obtained by applying the ASEM method to three representative clinical follow-up cases are illustrated in figure 5(c). Three colors were used to represent physiological uptake (in yellow), tumor response (in blue) area and stable tumor (in green), underlying the partial response status of these patients. The color intensity associated to the voxels classified as responders or stable is determined by the SUV relative variation ($\Delta$SUV) between the first and the second follow-up scans. The segmented maps of the pre-treatment and post-treatment scans, computed with the adaptive threshold and the ASEM methods are illustrated in figure 5(d–g).

The quantitative measurements estimated for each clinical case individually are shown in table 2, and mean measurements, estimated for all the patients are shown in table 3. The clinical cases were more challenging to analyze than the simulated cases, due to a combination of noisier and more heterogeneous background and tumor uptake distributions. Since the patients were classified as partial responder, the tumor uptakes were expected to exhibit a significant decrease of the $SUV_{max}$ (at least 30%) between the pre- and post-treatment scans. Similarly, the tumor volumes and associated mean tracer uptakes should also decrease (at least by 20 to 30%) and this can be confirmed visually.

**Table 2.** Measurements of volume, $SUV_{mean}$ and $SUV_{max}$ evolution, computed with the adaptive threshold and the ASEM methods for each patient.

| Patient | Method | $\Delta SUV_{mean}$ (%) | $\Delta V$ (%) | $\Delta SUV_{max}$ without PVC (%) | $\Delta SUV_{max}$ with PVC (%) |
|---|---|---|---|---|---|
| 1 | Tb 1 | −35.7 | −79.8 | −51.7 | −35.8 |
| | Tb 2 | −33.5 | −83.2 | | |
| | ASEM | −33.1 | −80.4 | | |
| 2 | Tb 1 | −69.5 | −86.3 | −74.2 | −76.9 |
| | Tb 2 | −74.2 | −64.9 | | |
| | ASEM | −71.4 | −65.2 | | |
| 3 | Tb 1 | −59.7 | −81.4 | −60.2 | −66.9 |
| | Tb 2 | −68.9 | −40.0 | | |
| | ASEM | −62.7 | −59.5 | | |
| 4 | Tb 1 | −27.0 | −66.1 | −32.7 | −25.7 |
| | Tb 2 | −24.2 | −71.7 | | |
| | ASEM | −26.7 | −59.2 | | |
| 5 | Tb 1 | −28.4 | −41.1 | −28.1 | −42.2 |
| | Tb 2 | −35.2 | −12.9 | | |
| | ASEM | −20.6 | −52.5 | | |
| 6 | Tb 1 | −87.3 | 402.5 | −56.1 | −87.8 |
| | Tb 2 | −85.7 | 74.1 | | |
| | ASEM | −83.3 | −81.9 | | |
| 7 | Tb 1 | −65.9 | 361.0 | −58.8 | −66.9 |
| | Tb 2 | −60.7 | −16.1 | | |
| | ASEM | −47.9 | −59.8 | | |

**Table 3.** Mean measurements of volume, $SUV_{mean}$ and $SUV_{max}$ evolution, computed with the adaptive threshold and the ASEM methods for all patients.

| Method | $\Delta SUV_{mean}$(%) | $\Delta V$ (%) | $\Delta SUV_{max}$ without PVC (%) | $\Delta SUV_{max}$ with PVC (%) |
|---|---|---|---|---|
| Tb 1 | $-53.3 \pm 23.2$ | $58.4 \pm 221.7$ | | |
| Tb 2 | $-54.7 \pm 23.6$ | $-30.7 \pm 53.5$ | $-51.7 \pm 16.2$ | $-57.5 \pm 23.0$ |
| ASEM | $-49.4 \pm 23.9$ | $-65.5 \pm 11.3$ | | |

With or without PVC, variations of $SUV_{max}$ were higher than 30% ($-52 \pm 16\%$ without PVC and $-58 \pm 23\%$ with PVC), as expected by their partial responder status. Regarding the other quantitative measurements, there was no significant ($p > 0.05$) difference between the mean variation of $SUV_{mean}$ obtained with each observer using the adaptive threshold ($-53 \pm 23$ for $Tb_1$, $-55 \pm 24\%$ for $Tb_2$). $SUV_{mean}$ variation deduced from the ASEM fusion maps was slightly, although not significantly ($p > 0.05$) lower ($-49 \pm 24\%$).

By contrast, the variations of tumor volumes were significantly different for all approaches. The variations measured by the two observers using the adaptive threshold were significantly ($p < 0.0001$) different ($58 \pm 222\%$ for $Tb_1$ and $-31 \pm 54\%$ for $Tb_2$). The results obtained with the ASEM method were also significantly different from both adaptive threshold results ($-66 \pm 11\%$). They were also much more homogeneous across the entire group of patients

(11% standard deviation only). These results can be explained by analyzing the quantitative parameters of each patient individually. Among the seven clinical cases, the adaptive threshold and ASEM performed differently. On the one hand, for patient 1, 2 and 4, the variation of $SUV_{mean}$ and tumor volume estimated with the adaptive threshold and the ASEM method were similar and pertinent with respect to the partial responder status of the patients. On the other hand, for patient 3 and 5, the tumor volume variation estimated by the two clinicians using adaptive thresholding were significantly different with a factor of 2 to 4 between the two measured variations ($-81$ and $-40\%$ for patient 3, $-41$ and $-13\%$ for patient 5). Finally, the use of adaptive threshold on patient 6 and 7 led to completely aberrant values (above +400 and +75%, and +360 and $-16\%$) contrary to the ASEM method that produced much more consistent volume variation results.

Different behavior of the adaptive threshold and ASEM method, three clinical follow-up cases corresponding to patients 2, 5 and 7 are illustrated in figure 5. Regarding patient 2, both adaptive thresholding applied independently to each PET scan and the proposed fusion method resulted in similar measurements leading to similar segmented maps. The first follow-up scan of patient 3 clearly exhibited a heterogeneous uptake within the tumor as shown in figure 5(a). The two segmentation maps obtained with $T_b$ clearly underestimated the overall tumor volume as shown in figure 5(d) and (f) by excluding the central part of the functional uptake, whereas ASEM included it. The poor reproducibility of the adaptive thresholding methodology is also emphasized for patient 7, for which the volume variation of the first observer is clearly overestimated (+360%) contrary to the second observer which tends to under estimate this volume evolution ($-16\%$) while the ASEM method estimated it at $-60\%$.

## 4. Discussion

The use multiple PET scans for response to therapy assessment is rising in oncology, due to the need to assess response to therapy earlier, in order to improve patient's management in radiotherapy and/or chemotherapy. On the other hand, the use of different radiotracers to visualize processes such as cellular proliferation or hypoxia for instance is generating a large amount of research especially in radiotherapy and early therapy assessment (Shields 2003, Vaupel and Mayer 2007).

The aim of this study was to propose a fusion method based on the multi-observation framework in order to specifically address the simultaneous analysis of multiple follow-up PET scans in the context of response to therapy assessment. The use of a fusion method taking into consideration both scans at the same time was expected to produce more reliable results than independent delineations performed on both scans separately. The ASEM method demonstrated the ability to merge patient follow-up PET scans through unsupervised Bayesian estimation, with especially good results on the first scan (error $-2.6 \pm 8\%$) and mostly good results on the second scan, with however a few cases that prove more difficult especially considering the second scan, therefore leading to a higher mean error and standard deviation of $-9 \pm 25\%$. On simulated datasets, the adaptive threshold applied independently on both images led to higher errors than the ASEM fusion with a systematic overestimation for both the first and second scan ($+28 \pm 17\%$ and $+30 \pm 15\%$, respectively). In the real clinical datasets however, a significantly higher variability in the quantitative parameters measured with the adaptive threshold method was observed for four patients out of seven.

These results can be explained by the fact that simulated data were generated with low and uniform physiological uptake and considering homogenous tumor uptake, as well as only one user to manually determine the background region of interest. However, noisy and heterogeneous uptake in nearby healthy tissues are very common in actual clinical datasets

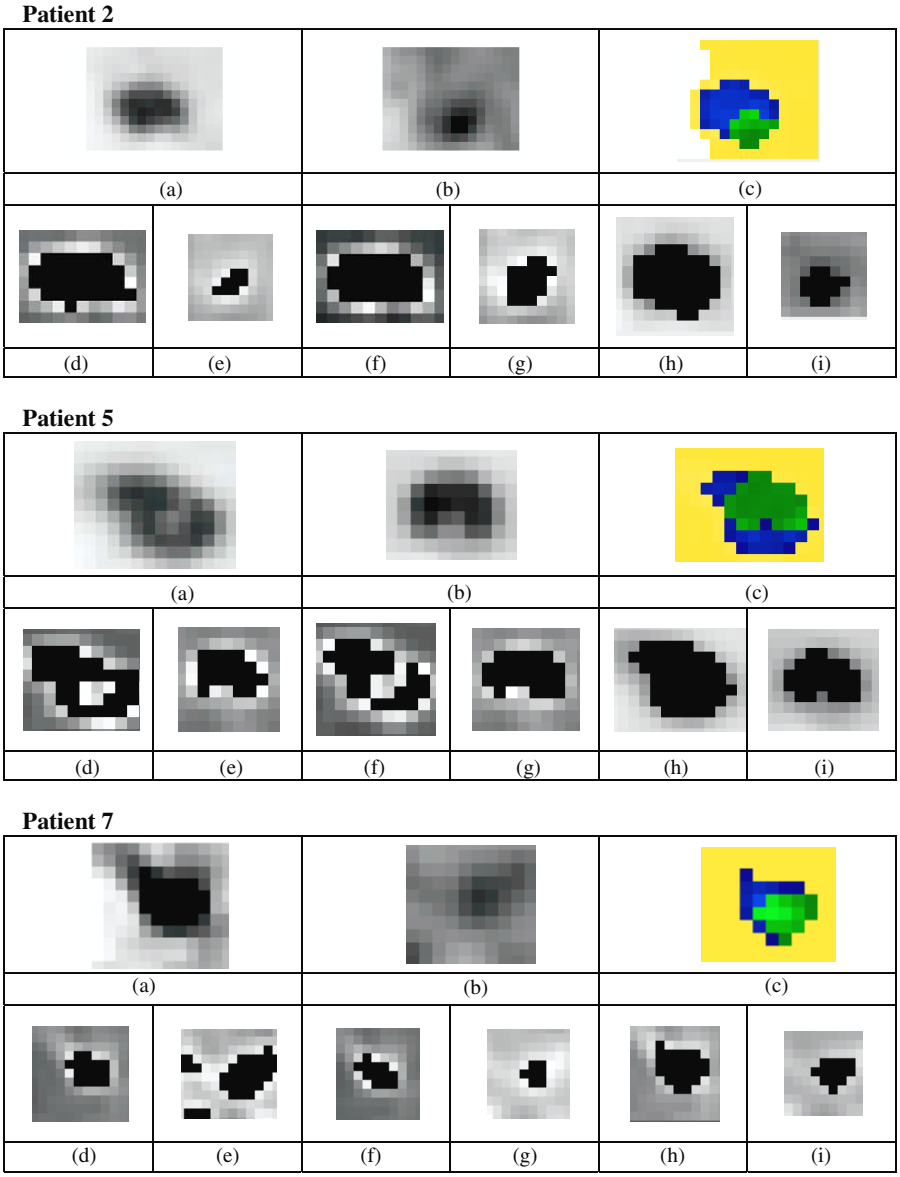



**Figure 5.** (a) and (b) Clinical follow-up tumors, (c) ASEM fusion maps, individual segmented map of the two clinicians (d), (e) and (f), (g) with the adaptive threshold, (h) and (i), the ASEM method, for three real clinical datasets.

affecting the delineation process, for instance in esophageal cancer for which the tumor is located in the mediastinum and close to lung tissues. ASEM seemed much more resilient with respect to non-uniform background and reduced tumor-to-background contrasts thanks to the use of spatial (the local adaptive priors) and inter-observation correlations (multi-observation framework using covariance matrices in the observation model) within the ASEM fusion. It is to be emphasized that the ASEM fusion requires accurate co-registration of PET datasets with a target registration error of 1 voxel (4 mm) as demonstrated by the results obtained by shifting the second scan in the simulated datasets by one and two voxels (4 and 8 mm) in random directions. The fusion was rather robust to 1 voxel spatial shifting; however, volume errors regarding the second scan became higher when shifting by 2 voxels was applied, which can be explained by the fact that as second scan tumor volumes are usually smaller, a misregistration can lead to a higher impact on the volume error of such small volumes.

Despite being more dependent on the noise and the images reconstruction, $SUV_{max}$ is nevertheless a parameter that is most commonly used in clinical routine to assess and quantify tumor evolution and response to therapy (Weber 2007, Nahmias and Wahl 2008). However, the variation of $SUV_{max}$ only may be not sufficient to characterize the tumor response, without taking into account the information of the metabolic tumor volume, especially for early therapy assessment. The $SUV_{mean}$ is considered more reproducible than $SUV_{max}$, but may depend on the definition of the tumor volume (Tylski *et al* 2010).

In this study, three quantitative indexes, namely $SUV_{max}$, $SUV_{mean}$ and volume were computed with both methodologies. An analysis of volume and $SUV_{mean}$ variations give additional features to characterize the tumor response. In our results, SUV measurements variations between pre- and post-treatment scans were similarly independent on the delineation used. Tumor volume variations measured by the clinicians using the adaptive method were close for patients 1, 2 and 4, and were significantly different for patients 3 and 5, emphasizing the user dependence of such method in the presence of heterogeneous physiological uptake, which is often the case for esophageal tumors (Hatt *et al* 2011). They were in addition aberrant for patients 6 and 7, demonstrating the accuracy limitation of such an approach. By contrast, tumor volume variations measured by ASEM were much more homogeneous across the group of patients ($-66 \pm 11\%$). The poor reproducibility of the adaptive method was first mentioned by Nestle *et al* (2005) in the case of non-small cell lung cancer. The measurements and segmented maps obtained with the ASEM method were more appropriate considering the known partial responder status of these patients. The combination of pertinent quantitative indexes such as the metabolic volume and activity concentration in the tumor, measured with a robust method could be valuable to thoroughly assess tumor response, as illustrated with the use of the ASEM method on the clinical datasets.

## 5. Conclusion

A fusion method based on the multi-observation Bayesian framework was proposed to assess multi-PET scans in the context of therapy response. Using the Bayesian framework, the proposed method can potentially be applied to patient follow-up and multi-tracer datasets in order to assess accurate treatment response and tumor volume definition by automatically delineating the different variations of activity within the tumor. In this study, the multi-observation method has been applied to simulated and clinical follow-up PET images and compared with current threshold-based methods used in clinical practice for assessment of the therapeutic response. On simulated datasets, the adaptive threshold applied independently on both images led to higher errors than the ASEM fusion. The adaptive threshold proves unreliable for more than half the patients, whereas ASEM produced measurements in line with

what could be expected with respect to the classification of the considered patients. Future work will also consider more than two PET scans within the context of therapy response assessment, as well as multi-tracer studies in order to adapt the proposed fusion approach for the definition of multi-tracer PET target volumes in radiotherapy, especially for dose boosting or dose painting scenarios in radiotherapy (Sovik *et al* 2009).

**Acknowledgments**

We would like to thank Professors M Allard and P Fernandez (Department of Nuclear Medicine, Bordeaux University Hospital) for useful discussions in the course of this study.

## Appendix

Let us consider two random processes $\mathbf{Y} = (\boldsymbol{y_t})_{t \leqslant T}$ and $\mathbf{X} = (x_t)_{t \leqslant T}$, modeling the observations and the fusion map, respectively. Considering a mixture of multi-dimensional Gaussian density probability functions, the distribution of $(\mathbf{X}, \mathbf{Y})$ is hence defined by the priors $\Pi_k$, the mean vectors $\boldsymbol{\mu_k}$ and covariance matrices $\Gamma_k$ associated to each of the $K$ classes in the mixture.

(1) The parameters $\left(\Pi_k^0, \boldsymbol{\mu}_k^0, \Gamma_k^0\right)$ defining the Gaussian mixture of the $(\mathbf{X}, \mathbf{Y})$ distribution and the class number $K$ are initialized with a FKM algorithm based on entropy criterion.

(2) The mixture parameters are then computed with the SEM algorithm by sampling several realizations of $\mathbf{X}$ according to its posterior distribution $p(\mathbf{X}|\mathbf{Y})$ defined for all the voxels $t \in [1, T]$ and each class $k \in [1, K]$ as

$$p(X_t = k|\boldsymbol{Y_t}) = \frac{\pi_{t,k} \times f(\boldsymbol{Y_t}, \boldsymbol{\mu_k}, \Gamma_k)}{\sum_{q=1}^{K} \pi_{t,q} \times f(\boldsymbol{Y_t}, \boldsymbol{\mu_q}, \Gamma_q)}, \tag{A.1}$$

where $\pi_{t,k}$ is the adaptive prior of the voxel $t$ and the class $k$ defined at the step 4, and $f(\boldsymbol{Y_t}, \boldsymbol{\mu_k}, \Gamma_k)$ is the multi-dimensional Gaussian defined for the $k$th class by the mean vectors $\boldsymbol{\mu_k}$ and covariance matrices $\Gamma_k$.

(3) For each voxel and associated observation vector $\boldsymbol{Y_t}$ with $t \in [1, T]$, a posterior realization called $R = (r_1 \ldots r_T)$ is sampled and a partition $Q = (Q_1, \ldots, Q_K)$ is defined as

$$Q_k = \{r_t | r_t = k\}, \tag{A.2}$$

where $Q_k$ is the partition associated to the $k$ class. Using these realizations, the parameters of the Gaussian mixture are estimated.

(4) First, in the adaptive framework, priors are re-estimated using a local neighboring 3D cube, hence priors for each voxel depend on its position in the image and the current state of its neighbors in the posterior realization. Replacing global prior $\Pi_k$, local priors $\pi_{t,k}$ are defined for each voxel and each class as

$$\pi_{t,k} = \frac{1}{\mathrm{Card}(C_t)} \sum_{j \in C_t} \delta(r_j, k) \quad \text{for} \quad k \in [1, K], \tag{A.3}$$

where $C_t$ is the estimation cube and $\delta$ is the Dirac function. For our application we chose a cube of size $(3 \times 3 \times 3)$ voxels.

(5) The mean vector associated to the $k$th class can be computed for each $b$ image with

$$\mu_k^{(b)} = \frac{\sum_{t \in Q_k} Y_t^{(b)}}{\mathrm{Card}(Q_k)}, \quad \text{for} \quad k \in [1, K], \quad \text{for} \quad b \in [1, B]. \tag{A.4}$$

(6) The covariance matrix associated to the $k$th class is defined as

$$\Gamma_k = \frac{\sum_{t \in Q_k} [\boldsymbol{Y}_t - \boldsymbol{\mu_k}]^t [\boldsymbol{Y}_t - \boldsymbol{\mu_k}]}{\mathrm{Card}(Q_k)}, \quad \text{for} \quad k \in [1, K]. \tag{A.5}$$

(7) The decision step based on the maximum likelihood criteria computes the posterior probability $p(\mathbf{X}|\mathbf{Y})$ and selects for each voxel the class that maximizes it

$$k_{\max} = \arg \max_{k \in [1,\ldots,K]} p\left(X_t = k | \boldsymbol{Y}_t\right), \quad \forall t, \tag{A.6}$$

where $K_{\max}$ is the estimated maximized class.

## References


Bentzen S M 2005 Theragnostic imaging for radiation oncology: dose-painting by numbers *Lancet Oncol.* **6** 112–7

Boussion N, Cheze Le Rest C, Hatt M and Visvikis D 2008 Incorporation of wavelet based denoising in iterative deconvolution for partial volume correction in whole body PET imaging *Eur. J. Nucl. Med. Mol. Imaging* **36** 1064–75

Celeux G and Diebolt J 1986 L'algorithme SEM: un algorithme d'apprentissage probabiliste pour la reconnaissance de mélanges de densités *Revue de statistique appliquée* **34** 35–52

Delignon Y, Marzouki A and Pieczynski W 1997 Estimation of generalized mixtures and its application in image segmentation *IEEE Trans. Image Process.* **6** 1364–75

El Naqa I *et al* 2009 Exploring feature based approaches in PET images for predicting cancer treatment outcomes *Pattern Recognit.* **42** 1162–71 190

Erdi Y E, Mawlawi O, Larson S M, Imbriaco M, Yeung H, Finn R and Humm J L 1997 Segmentation of lung lesion volume by adaptative positron emission tomography image thresholding *Cancer* **80** 2505–9

Hatt M, Cheze Le Rest C, Aboagye E O, Kenny L M, Rosso L, Turkheimer F E, Albarghach N M, Pradier O and Visvikis D 2010 Reproducibility of 18F-FDG and 18F-FLT PET tumour volume measurements *J. Nucl. Med.* **51** 1368–76

Hatt M, Cheze Le Rest C, Pradier O and Visvikis D 2009 Automatic PET tumour delineation for patient's follow-up and therapy assessment *J. Nucl. Med.* **50** 182

Hatt M, Visvikis D, Albarghach N M, Tixier F, Pradier O and Cheze-le Rest C 2011 Prognostic value of 18F-FDG PET image-based parameters in œsophageal cancer and impact of tumour delineation methodology *Eur. J. Nucl. Med. Mol. Imaging* **38** 1191–202

Jarritt H, Carson K, Hounsel A R and Visvikis D 2006 The role of PET/CT scanning in radiotherapy planning *Br. J. Radiol.* **79** S27–35

Krak N C *et al* 2005 Effects of ROI definition and reconstruction method on quantitative outcome and applicability in a response monitoring trial *Eur. J. Nucl. Med. Mol. Imaging* **32** 294–301

Krishnapuram R and Keller J M 1994 Fuzzy and possibilistic clustering methods for computer vision *SPIE Inst. Ser.* **12** 133–59

Larson S *et al* 1999 Tumour treatment response based on visual and quantitative changes in globals using PET-FDG imaging, the visual response score and the change in total lesion glycolysis *Clin. Positron Imaging* **2** 159–71 160

Le Maitre A, Segars W P, Marache S, Reilhac A, Hatt M, Tomei S, Lartizien C and Visvikis D 2009 Incorporating patient specific variability in the simulation of realistic whole body 18F-FDG distributions for oncology applications *Proc. IEEE* **97** 2026–38

Le Pogam A, Descourt P, Hatt M, Boussion N and Visvikis D 2009 A combined 3-D wavelet and curvelet approach for edge preserving denoising in emission tomography *J. Nucl. Med.* **50** 533

Lin C *et al* 2007 Early $^{18}$F-FDG PET for prediction of prognosis in patient with diffuse large B-cell lymphoma: SUV-based assessment versus visual analysis *J. Nucl. Med.* **48** 1626–32

Mankoff D A, Muzi M and Krohn K A 2003 Quantitative PET imaging to measure tumour response to therapy: what is the best method? *Mol. Imaging Biol.* **5** 281–5

Masson P and Pieczynski W 1993 Adaptive mixture estimation and unsupervised local Bayesian image segmentation *IEEE Trans. Geosci. Remote Sens.* **31** 618–33

Nahmias C and Wahl L 2008 Reproducibility of standardized uptake value measurements determined by 18F-FDG PET in malignant tumours *J. Nucl. Med.* **49** 1804–8 158 159

Necib H, Dusart M, Tylski P, Vanderlinden B and Buvat I 2008 Detection the tumour changes between two FDG PET scans using parametric imaging *J. Nucl. Med. Meeting Abstracts* **49** 121 161

Nestle U, Kremp S, Schaefer-Schuler A, Sebastian-Welch C, Hellwig D, Rübe C and Kirsch C M 2005 Comparison of different methods for delineation of 18F-FDG PET-positive tissue for target volume definition in radiotherapy of patients with non-small cell lung cancer *J. Nucl. Med.* **46** 1342–8

Ouksili Z *et al* 2007 Accurate PET/PET registration of serial to assess lung tumour evolution *4th IEEE Int. Symp. on Biomedical Imaging* pp 732–5

Peng A and Pieczynski W 1995 Adaptive mixture estimation and unsupervised local Bayesian image segmentation *Graph. Models Image Process.* **57** 389–99

Pieczynski W 2003 Modèles de Markov en traitement d'images *Traitement du Signal* **20** 255–77

Provost J N 2001 Classification bathymétrique en imagerie multispectrale SPOT *PhD Thesis* Université Bretagne Occidentale

Shields A F 2003 PET imaging with 18F-FLT and thymidine analogs: promise and pitfalls *J. Nucl. Med.* **44** 1432–4

Soret M, Bacharach S L and Buvat I 2007 Partial-volume effect in PET tumor imaging *J. Nucl. Med.* **48** 932–45

Sovik A, Malinen E and Olsen D R 2009 Strategies for biologic dose escalation: a review *Int. J. Radiat. Oncol. Biol. Phys.* **73** 650–8

Therasse P *et al* 2000 New guidelines to evaluate the response to treatment in solid tumors *J. Natl Cancer Inst.* **92** 205–16

Tylski P, Stute S, Grotus N, Doyeux K, Hapdey S, Gardin I, Vanderlinden B and Buvat I 2010 Comparative assessment of methods for estimating tumor volume and standardized uptake value in FDG PET *J. Nucl. Med.* **51** 268–76

Vaupel P and Mayer A 2007 Hypoxia in cancer: significance and impact on clinical outcome *Cancer Metastasis Rev.* **26** 225–39

Wahl R, Jacene H, Kasamon Y and Lodge M 2009 From RECIST to PERCIST: evolving considerations for PET response criteria in solid tumors *J. Nucl. Med.* **50** 122–50

Weber W 2007 $^{18}$F-FDG PET in non-Hodgkin's lymphoma: qualitative or quantitative? *J. Nucl. Med.* **48** 1580–2

Weber W, Ziegler S, Thodtmann R, Hanauske A and Schwaiger M 1999 Reproducibility of metabolic measurements in malignant tumours using FDG PET *J. Nucl. Med.* **40** 1771–7 158 159

World Health Organization 1979 *Handbook for Reporting Results of Cancer Treatment* (USA: World Health Organization)

Zaidi H and El Naqa I 2010 PET-guided delineation of radiation therapy treatment volumes: a survey of image segmentation techniques *Eur. J. Nucl. Med. Mol. Imaging* **37** 2165–87

Cancer/Radiothérapie xxx (2011) xxx–xxx

Disponible en ligne sur
SciVerse ScienceDirect
www.sciencedirect.com

Elsevier Masson France
EM|consulte
www.em-consulte.com

Revue générale

# Méthodologies de définition automatique des volumes métaboliquement actifs en TEP : évaluation et perspectives

## *Metabolically active volumes automatic delineation methodologies in PET imaging: Review and perspectives*

M. Hatt[a,*], N. Boussion[a,b], C. Cheze-Le Rest[a], D. Visvikis[a], O. Pradier[a,b]

[a] *Inserm U650 LaTIM, CHU Morvan, 5, avenue Foch, 29609 Brest, France*
[b] *Département de radiothérapie, CHRU Morvan, 5, avenue Foch, 29609 Brest, France*

INFO ARTICLE

*Historique de l'article :*
Reçu le 22 avril 2011
Reçu sous la forme révisée le 31 mai 2011
Accepté le 4 juillet 2011
Disponible sur Internet le xxx

*Mots clés :*
TEP
Segmentation
Contours
Radiothérapie
Suivi thérapeutique
Validation

RÉSUMÉ

La tomographie par émission de positons (TEP) est dorénavant un outil de référence en routine clinique en oncologie, notamment dans le cadre du diagnostic. Des applications plus récentes, telles la prise en charge et le suivi thérapeutique, ou la définition de cibles en radiothérapie nécessitent une détermination rapide, précise et robuste des volumes métaboliquement actifs sur les images d'émission, ce qui ne peut être obtenu par contours manuels. Ce besoin clinique a motivé de nombreux développements ces dernières années pour la mise au point de méthodes automatiques. Cette revue propose une vue d'ensemble de ces méthodologies, et discute leur mise en œuvre et leur validation méthodologique et/ou cliniques. Des perspectives sur les travaux encore à accomplir sont également suggérées.



*Keywords:*
PET
Segmentation
Contouring
Radiotherapy
Therapy follow-up
Validation

ABSTRACT

PET imaging is now considered a gold standard tool in clinical oncology, especially for diagnosis purposes. More recent applications such as therapy follow-up or tumor targeting in radiotherapy require a fast, accurate and robust metabolically active tumor volumes delineation on emission images, which cannot be obtained through manual contouring. This clinical need has sprung a large number of methodological developments regarding automatic methods to define tumor volumes on PET images. This paper reviews most of the methodologies that have been recently proposed and discusses their framework and methodological and/or clinical validation. Perspectives regarding the future work to be done are also suggested.



## 1. Introduction

La tomographie par émission de positons (TEP) est dorénavant un outil de référence en routine clinique en oncologie, notamment pour le diagnostic [1]. Des applications plus récentes de cette imagerie fonctionnelle concernent la prise en charge et le suivi thérapeutique, ou l'identification et la définition des cibles en radiothérapie [2,3]. Ces dernières applications sont encore au stade du développement. Elles nécessitent des efforts en termes de standardisation [4], notamment multicentriques, ainsi que la mise au point d'outils permettant une quantification robuste, précise et reproductible. En particulier, la mise à disposition de méthodes permettant la définition automatique, rapide et fiable des volumes métaboliquement actifs (VMA) des tumeurs a été identifiée comme un besoin important et urgent [2,5]. Les applications visées sont notamment la mise au point de nouveaux indices

* Auteur correspondant.
*Adresse e-mail :* hatt@univ-brest.fr (M. Hatt).

1278-3218/$ – see front matter © 2011 Société française de radiothérapie oncologique (SFRO). Publié par Elsevier Masson SAS. Tous droits réservés.
doi:10.1016/j.canrad.2011.07.243

Pour citer cet article : Hatt M, et al. Méthodologies de définition automatique des volumes métaboliquement actifs en TEP : évaluation et perspectives. Cancer Radiother (2011), doi:10.1016/j.canrad.2011.07.243

pronostiques ou prédictifs de la réponse, le suivi thérapeutique et l'évaluation de la réponse et la définition des contours des volumes tumoraux macroscopiques (*gross tumour volumes* [GTV]) en radiothérapie [2,6–10]. Ce besoin a notamment émergé suite aux efforts déployés depuis le début des années 2000 pour tenter de mettre en œuvre le « volume cible biologique », avec l'utilisation de radiotraceurs différents et plus spécifiques que le ($^{18}$F)-fluorodésoxyglucose [($^{18}$F)-FDG], dont l'utilisation reste largement majoritaire en oncologie. Les traceurs permettant de visualiser la prolifération cellulaire [($^{18}$F)-fluorothymidine (FLT)] ou l'hypoxie [($^{64}$Cu)-diacétyl-bis(N4-méthylthiosemicarbazone (ATSM)], ($^{18}$F)-fluoromisonidazole (FMiso)] suscitent un intérêt particulier en radiothérapie car ils permettent d'identifier des régions plus agressives ou radiorésistantes. Ils prennent une importance accrue en combinaison avec les nouvelles techniques d'irradiation ciblée comme la radiothérapie conformationnelle avec modulation d'intensité (RCMI) et les scénarii d'escalade de dose [11] et de *dose painting* [12]. Toutefois, la méthode standard de définition des volumes tumoraux macroscopiques reste le contour manuel coupe par coupe par le radiothérapeute sur les volumes obtenues par tomodensitométrie (TDM) de planification. Cette dernière bénéficie d'une résolution spatiale élevée de l'ordre du millimètre, supérieure à celle offerte par la TEP (de l'ordre de 4 à 6 mm). Plusieurs travaux ont étudié l'impact de l'utilisation des images fusionnées TEP/tomodensitométrie sur la reproductibilité des contours ainsi que sur la taille et la forme des volumes tumoraux macroscopiques. Par rapport à l'utilisation de l'image anatomique seule, les contours réalisés sur la fusion TEP/tomodensitométrie mènent à des volumes tumoraux macroscopiques en général plus reproductibles et souvent significativement plus petits ou plus grands en fonction des localisations et des cas [13–16]. Actuellement, la TEP est utilisée en routine clinique au mieux comme un guide visuel à la définition des volumes tumoraux macroscopiques. La planification réellement fondée sur l'imagerie fonctionnelle (avec le traceur FDG, seul ou combinant différents traceurs), bien qu'ayant été suggérée [17], est loin d'être une réalité clinique. Une des limites majeures à la réalisation d'une telle méthodologie est sa complexité : si la variabilité inter- et intra-utilisateur de la détermination manuelle des volumes tumoraux macroscopiques sur l'imagerie anatomique peut être significative [14], celle constatée sur l'imagerie fonctionnelle est plus importante et incompatible avec une pratique clinique [18].

Nous proposons dans cette revue une évaluation critique des solutions qui ont été proposées jusqu'à présent pour assister le clinicien dans la démarche chronophage et complexe de définition des contours du volume métaboliquement actif sur les images de TEP. Nous discuterons des manques actuels qu'il faut combler pour parvenir à des solutions pouvant être mises en pratique en routine clinique.

## 2. Évaluation critique

### 2.1. Problématique des méthodes fondées sur des seuillages

Les premiers travaux visant à établir une définition des volumes métaboliquement actifs en TEP sont fondés majoritairement sur des méthodes de seuillage déterministe (fixe ou adaptatif) des valeurs des voxels de l'image. Une liste détaillée de ces travaux peut être consultée dans la revue de Dewalle et al. [19]. Le cas de la TEP est particulier dans le sens où les seuillages ont pris une grande importance en termes de popularité auprès des cliniciens et en conséquence un grand nombre de publications les utilisant. Le principal attrait de ces approches est leur simplicité méthodologique et leur facilité d'utilisation. Elles reposent toutefois sur des hypothèses simplificatrices, en particulier concernant la distribution du radiotraceur au sein du volume métaboliquement actif. Cela explique la grande variabilité de valeurs de seuils indiquées dans les publications (de 20 à 75 % du *standard uptake value* [SUV] maximum) et leur manque de précision et de robustesse [18,20–24]. Afin d'apporter une réponse au problème fondamental de l'utilisation d'une valeur fixe de seuil, il a été proposé d'adapter cette valeur aux caractéristiques du volume métaboliquement actif et de son environnement immédiat. Ces « seuils adaptatifs » se fondent sur une estimation approximative du contraste entre le volume métaboliquement actif et le « fond » physiologique, en lien avec une optimisation propre au système et au protocole d'acquisition. Les résultats sont ainsi améliorés, mais ces approches souffrent encore de nombreuses limites. L'estimation initiale du contraste est le plus souvent dépendante de l'utilisateur [18,22], et l'optimisation nécessaire rend la méthode spécifique au système et au protocole du centre clinique. Une variabilité significative peut en effet être observée entre des centres cliniques différents utilisant pourtant le même modèle de scanner [25]. En l'absence de standardisation, ces approches peuvent difficilement être utilisées dans le cadre d'études multicentriques. Ces techniques sont en outre susceptibles d'échouer dans des situations complexes, notamment de distribution hétérogène du radiotraceur [6,23,26], comme illustré en Fig. 1. Il est dorénavant acquis que les seuillages fixes sont à proscrire, et qu'en absence d'alternative, une segmentation manuelle est préférable [20]. Les seuillages adaptatifs sont susceptibles d'apporter une précision suffisante dans les cas simples, mais sont dépendants des utilisateurs et du système. Ils font globalement trop d'hypothèses simplificatrices pour pouvoir être considérés comme une solution d'avenir.

## 3. Méthodologies de segmentation d'images

Étant données les limites des méthodes à base de seuillage, il est pertinent de s'intéresser à des méthodologies fondées sur des approches de segmentation d'image plus avancées. Ces dernières ont en commun des outils d'analyse et de segmentation d'image ayant été développés auparavant dans d'autres domaines. Elles se distinguent par le type de méthodologie considérée, les éventuelles modifications apportées, leur niveau d'automatisation, la nécessité de pré- et/ou post-traitement(s) et par leur niveau de validation.

Bien que les spécificités propres à chaque localisation et type de tumeurs ou au radiotraceur considéré doivent être gardées à l'esprit, la définition automatique d'un volume métaboliquement actif sur une image de TEP reste avant tout une problématique de segmentation d'image. C'est dans cette optique que nous orientons cette revue critique. Certaines méthodologies ont été développées pour résoudre la problématique de façon globale, tandis que d'autres l'ont été pour une localisation particulière (sphère ORL, poumon, rectum...) ou une problématique plus spécifique encore (imagerie dynamique, contours en radiothérapie...). Notons également que l'immense majorité des travaux ont étudié le problème en considérant le ($^{18}$F)-FDG.

### 3.1. Méthodologies considérées

Il existe littéralement des centaines de méthodes de segmentation d'images, et plusieurs des plus connues ont été considérées par différents groupes pour étudier la problématique de la segmentation d'images de TEP afin de définir les volumes métaboliquement actifs. Certaines cherchent à identifier les contours dans l'image, d'autres tentent d'identifier les régions ou regroupements de voxels. Cela peut être fait suivant des critères variés comme les valeurs, les formes ou les textures, par des approches déterministes, statistiques et probabilistes, ou d'intelligence artificielle.

Pour citer cet article : Hatt M, et al. Méthodologies de définition automatique des volumes métaboliquement actifs en TEP : évaluation et perspectives. Cancer Radiother (2011), doi:10.1016/j.canrad.2011.07.243

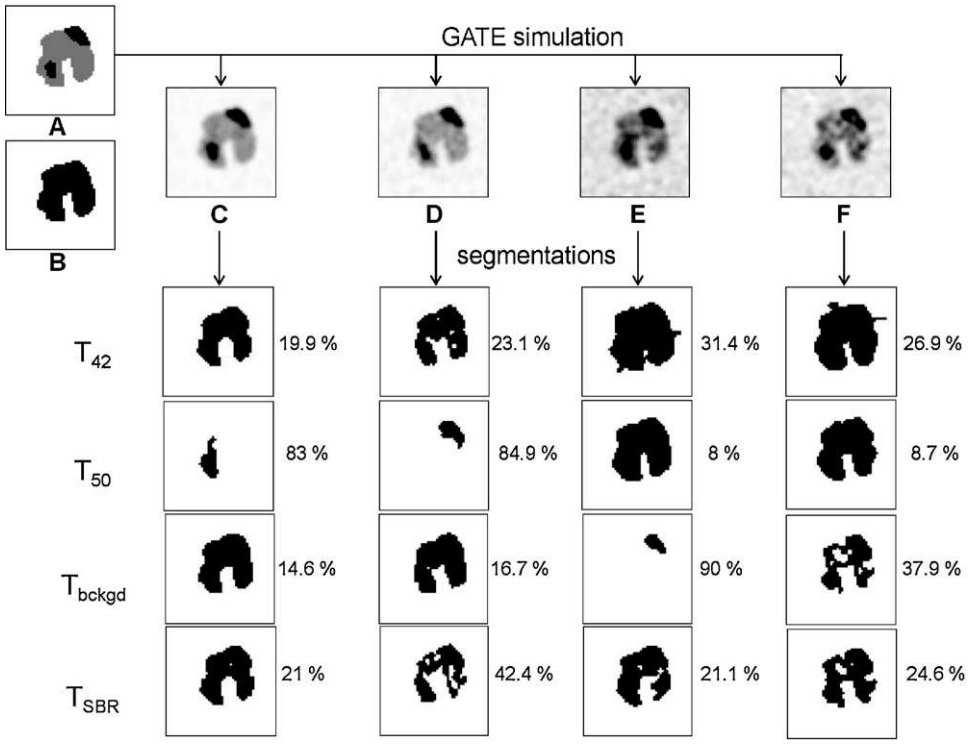



**Fig. 1.** Erreurs de classification (%) obtenues par des seuillages sur simulations Geant4 Application for Tomography Emission (GATE) [59]. A : vérité terrain simulée et B : binarisée pour le calcul d'erreurs ; C–F : simulations avec différents niveaux de contraste et bruit. $T_{42}$ et $T_{50}$ : seuils fixes à 42 et 50 % du maximum. $T_{bckgd}$ et $T_{SBR}$ : seuillages adaptatifs de Nestle et Daisne [19].

Une approche classique en segmentation d'images consiste en la détection de contours par analyse du gradient de l'image, afin de définir une région interne et une région externe au contour détecté. L'application directe de ce type de méthode en TEP est difficile car les contours sont flous et les gradients difficiles à identifier. Plusieurs études ont étudié l'adaptation de ces méthodes pour la définition de volume métaboliquement actif en TEP. Par exemple, la méthode du « partage des eaux » (*watershed*), qui est fondée sur l'analogie avec une surface topologique se remplissant d'eau [27]. Une autre approche est fondée sur la détection des pics de gradient pour identifier les contours des objets d'intérêt [28–32]. Il a également été proposé d'améliorer le résultat d'un seuillage adaptatif par l'utilisation d'un contour actif [33]. Notons qu'il a été suggéré également de faire évoluer de tels contours en prenant en compte les images de TEP et de tomodensitométrie simultanément [34].

Les méthodologies fondées sur le regroupement (*clustering*) flou non supervisé de voxels (en anglais *: Fuzzy C-Means* [FCM]) ont été utilisées par plusieurs groupes [21,35,36]. Notons que l'algorithme original est relativement simpliste et ne prend pas en compte les corrélations spatiales entre voxels, par exemple, et mène à des résultants décevants [21,36]. Il semble toutefois a priori assez bien adapté aux images de TEP, du fait de sa modélisation floue. Le groupe ayant proposé la méthode la plus aboutie fondée sur le FCM, a utilisé une version modifiée incorporant des informations supplémentaires comme la détection automatique du nombre de classes (ou *clusters*), la corrélation spatiale des voxels et l'analyse de l'hétérogénéité du traceur [36].

Le regroupement des voxels pour définir les régions tumorales et physiologiques peut aussi se faire par la différentiation statistique des valeurs dans l'image. Ce type de méthode est très utilisé en traitement d'images et a de nombreuses applications en imagerie satellite et astronomique notamment [37]. Le principe est de distinguer les voxels appartenant aux tumeurs de ceux appartenant aux tissus sains par leurs propriétés statistiques respectives. Elles sont toutefois assez peu adaptées au traitement d'images de TEP à cause de la difficulté de prise en compte du flou. Une approche simple consiste à faire croître la région de l'image correspondant au volume métaboliquement actif à partir du voxel d'intensité maximale, avec comme critère de croissance la moyenne et la variance [38]. Une autre approche est appliquée aux projections *maximum intensity projection* (MIP) pour bénéficier du contraste ainsi augmenté, et fait correspondre le volume métaboliquement actif défini sur la TEP d'origine grâce aux ensembles flous associés à un opérateur de fusion [35]. L'utilisation de mélanges de gaussiennes pour opérer la classification des voxels a été proposée également. Cette méthode ne considère pas de modèle spatial et nécessite une complexe estimation du nombre total de gaussiennes à utiliser et de celles qui sont à associer au volume métaboliquement actif [39]. Dans ce contexte, il est possible de modéliser l'information spatiale dans l'image par des modèles de Markov [40]. Il est aussi possible de modifier la modélisation des données au sein de ces outils, pour prendre en compte le flou et ainsi obtenir des résultats améliorés [41]. L'approche *fuzzy locally adaptive bayesian* (FLAB) utilise ce principe [18,21,22,26].

L'analyse de texture a également fait l'objet d'adaptation pour classifier les voxels d'images de TEP/tomodensitométrie [42,43]. Cette approche consiste à apprendre à l'algorithme, via une base de données d'images pour lesquelles la vérité est identifiée par des médecins, en quoi les voxels de tumeurs forment des textures différentes de celles des voxels physiologiques. Cet apprentissage permet la construction de classifieurs (par exemple des arbres de décision) servant à classifier les voxels d'une nouvelle image proposée en entrée de l'algorithme. Une méthode similaire dans son approche (apprentissage sur base de données pour générer un classifieur), mais utilisant des réseaux de neurones, a récemment été proposée [44].

D'autres approches peuvent être citées, notamment une étude qui s'est spécifiquement intéressée à la segmentation de tumeurs

Pour citer cet article : Hatt M, et al. Méthodologies de définition automatique des volumes métaboliquement actifs en TEP : évaluation et perspectives. Cancer Radiother (2011), doi:10.1016/j.canrad.2011.07.243

en détectant les fixations physiologiques situées à proximité pour éviter de les incorporer au volume métaboliquement actif [45]. Il a aussi été proposé de classifier les voxels comme tumoraux ou physiologiques en les regroupant en fonction de la courbe temps–activité (*time–activity curve*) les caractérisant [46]. Cette méthode ne peut toutefois être utilisée que sur des images paramétriques issues d'acquisitions de TEP dynamiques.

### 3.2. Prise en compte de la nature des images de tomographie par émission de positons

Des méthodes reconnues comme peu robustes au bruit et fondées sur des décisions de segmentation peu flexibles paraissent a priori mal adaptées à l'analyse d'images de TEP. C'est pourtant le cas des seuillages. Certaines approches, au contraire, proposent de modéliser explicitement les caractéristiques des images de TEP. C'est notamment le cas de l'approche FLAB [21]. D'autres méthodes ont modifié des algorithmes connus pour obtenir une précision satisfaisante. C'est le cas de l'algorithme par partage des eaux, pour lequel le contour d'origine a été modifié [27]. De façon similaire, l'algorithme FCM a été modifié pour incorporer l'information de corrélation spatiale et celle de l'hétérogénéité du traceur [36]. Les approches fondées sur des techniques issues de l'intelligence artificielle [42–44] ne proposent pas de modélisation explicite des caractéristiques des images de TEP, toutefois leur nature intrinsèque doit leur permettre d'apprendre ces caractéristiques pour s'y adapter. Une alternative consiste à améliorer la qualité des images de TEP préalablement à la segmentation. Ainsi, la méthodologie fondée sur la théorie des possibilités est appliquée aux projections MIP et non aux images d'origine, dans le but d'opérer la segmentation sur un contraste plus élevé [35]. Les approches utilisant l'analyse du gradient sont souvent associées à une ou plusieurs étapes de prétraitement visant à permettre une détection moins complexe des contours [32,34]. Ces prétraitements atténuent le bruit et réduisent le flou. La même méthodologie proposée par l'industrie MIMVista semble ne pas faire appel à de tels prétraitements[1] [28–31]. Cela pose la question de la dépendance à cette étape, car ces corrections sont délicates à étendre à l'ensemble des images de TEP [47,48]. D'autres approches ont par ailleurs démontré de bonnes performances sur les images d'origine sans nécessiter de tels prétraitement(s) [21,36].

Le développement d'une approche de segmentation automatique d'images de TEP nécessite, en effet, la prise en compte de plusieurs paramètres.

D'une part, ceux qui sont liés à la tumeur ou à l'organe à définir :

- hétérogénéité de la distribution du traceur au sein de la tumeur/organe ;
- hétérogénéité de la distribution du traceur au sein du fond physiologique ;
- contraste(s) mis en jeu entre l'objet et le fond, et au sein même de l'objet et du fond ;
- complexité de la forme de l'objet.

D'autre part, ceux qui sont liés à la nature de l'acquisition :

- niveau de bruit ;
- taille des voxels et échantillonnage spatial ;
- type et modèle du scanographe ;
- algorithme de reconstruction et ses paramètres.

[1] Il est important de souligner que les détails de la méthodologie ne sont pas disponibles, seuls les résultats étant présentés dans des papiers de conférences et non des articles en journaux à comité de lecture.

Les paramètres ayant le plus d'impact sur la précision et la robustesse d'une segmentation sont l'hétérogénéité de la distribution du traceur et les contrastes mis en jeu (Fig. 1). Les volumes métaboliquement actifs réels sont rarement sphériques et homogènes sur un fond homogène comme le supposent de nombreuses méthodes, et comme sont conçues les données de validation utilisées dans de nombreuses publications.

Les autres paramètres peuvent également avoir un impact important sur la qualité globale de l'image. La taille des voxels, par exemple, a un impact significatif sur la définition d'un contour. Plus les voxels utilisés pour définir la grille de reconstruction sont petits, meilleur est l'échantillonnage spatial des structures, et plus le contour peut être finement défini. Toutefois, la réduction de la taille des voxels dans la grille de reconstruction entraîne une baisse significative des statistiques disponibles en chaque voxel, et une augmentation significative du bruit (Fig. 2). Une alternative pour améliorer l'échantillonnage spatial consiste à suréchantillonner l'image, ce qui est d'ailleurs réalisé dans les stations commerciales pour la visualisation. Cela a toutefois un impact important sur l'aspect visuel des images et peut varier en fonction de la méthode utilisée (Fig. 3). Étant donnée la grande variabilité des modèles et types de scanner en activité, le manque de standardisation des protocoles d'acquisition, et le grand nombre d'algorithmes de reconstruction existant (et la possibilité de varier leurs paramètres), les images de TEP peuvent présenter des aspects très variés (Fig. 4). Se pose alors le problème de l'universalité et de la robustesse des méthodes proposées qui n'ont été validées que sur un nombre restreint de données. Enfin, les mouvements physiologiques, en particulier la respiration, ont un impact important, notamment pour les localisations thoraciques. Plusieurs études ont récemment concerné cette problématique [49,50], mais la grande majorité des méthodes ne prennent pas en compte ce paramètre dans leur validation. Soulignons toutefois qu'il n'est pas déraisonnable de faire l'hypothèse que les effets de la respiration puissent être corrigés en amont au cours ou après la reconstruction [51].

### 3.3. Validation

Afin de démontrer rigoureusement les performances d'une méthodologie, il faut évaluer :

- la précision absolue ;
- la robustesse ;
- la reproductibilité et la dépendance éventuelle à l'utilisateur.

La précision désigne la capacité de l'algorithme à définir la position, la forme et le volume du volume métaboliquement actif et pour l'évaluer, il est nécessaire de disposer d'une vérité terrain. Certains travaux ont suggéré l'utilisation des volumes segmentés sur les images tomodensitométriques comme vérité terrain, ce qui est aberrant car rien ne garanti que les volumes anatomiques et fonctionnels soient parfaitement superposés. La définition manuelle par un expert ne permet pas non plus de générer une vérité terrain satisfaisante du fait de la variabilité inter- comme intra-utilisateurs [14,18]. Une alternative est de définir un consensus de nombreux utilisateurs. Cela reste toutefois dépendant des utilisateurs, ainsi que de la façon dont le consensus est défini [52]. Dans le cas des images cliniques, c'est toutefois la seule alternative à l'analyse histopathologique qui consiste à extraire chirurgicalement l'objet et d'en réaliser des mesures macroscopiques. Cette opération comporte de nombreuses sources d'erreurs et d'approximations, l'objet pouvant subir des déformations, devant être coupé en tranches (ce qui peut provoquer des pertes de matériau), et les régions correspondant aux zones tumorales étant définies manuellement sur les coupes. Enfin, les données doivent être recalées spatialement,

Pour citer cet article : Hatt M, et al. Méthodologies de définition automatique des volumes métaboliquement actifs en TEP : évaluation et perspectives. Cancer Radiother (2011), doi:10.1016/j.canrad.2011.07.243

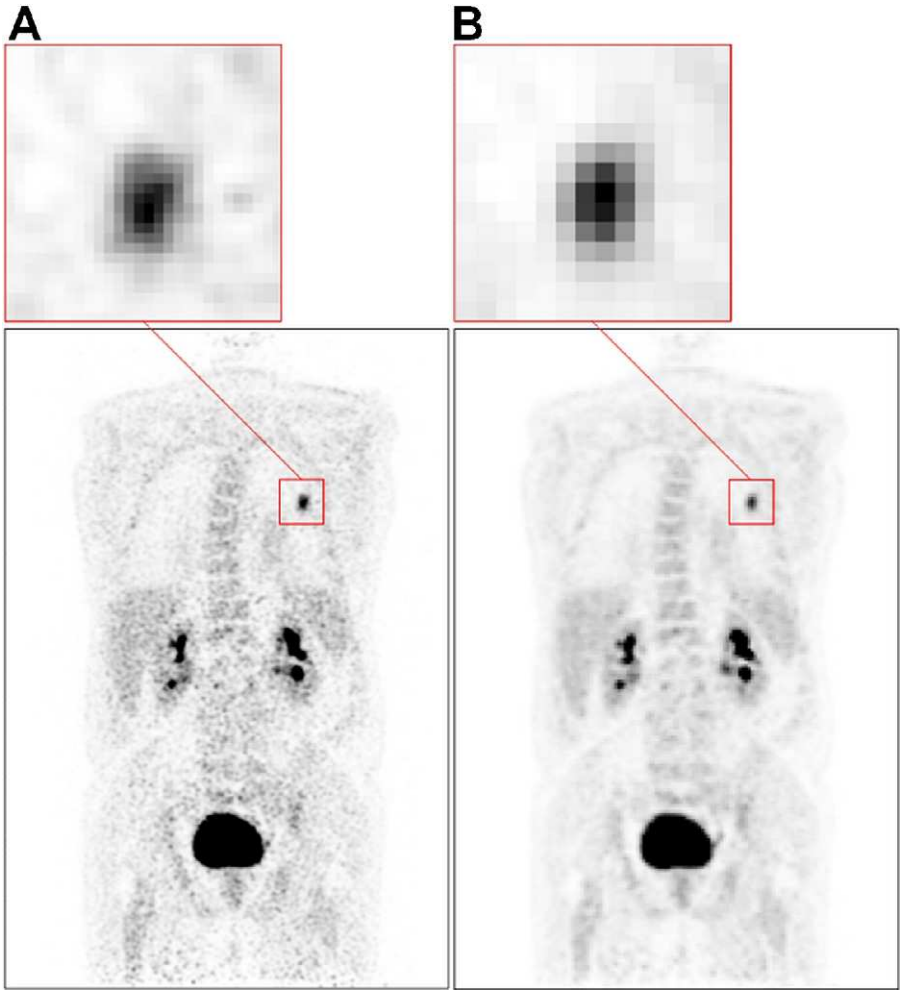



**Fig. 2.** Acquisition de tomographie par émission de positons (TEP) corps entier au ($^{18}$F)-FDG reconstruite avec des voxels de A : $2 \times 2 \times 2\,mm^3$ et B : $4 \times 4 \times 4\,mm^3$.

ce qui est complexe étant donnée la nature très différente des images. Peu de jeux de données existent [14,53–57]. Ils sont de taille réduite, de moins d'une dizaine à une trentaine de patients. Dans la majorité des cas, seul le diamètre maximal est mesuré, ce qui constitue une mesure réductrice du volume métaboliquement actif.

Une alternative à l'utilisation de données cliniques pour évaluer la précision consiste à réaliser des acquisitions de modèles physiques réels dont les dimensions sont connues, tels les fantômes NEMA IEC contenant des sphères de taille variable. La seule mesure pertinente qu'il est possible d'évaluer sur de telles données est la robustesse par rapport à des structures de tailles et de contrastes variables. En effet, les objets en question sont des sphères dont l'activité est homogène, placées sur un fond homogène, ce qui est simpliste par rapport aux volumes métaboliquement actifs réels. Des fantômes physiques permettant de générer des activités hétérogènes et/ou des formes plus complexes ont été proposées [40], mais restent loin de la complexité des images cliniques réelles et sont complexes à produire et utiliser. Une étude rigoureuse et complète de la robustesse devrait idéalement être réalisée sur des

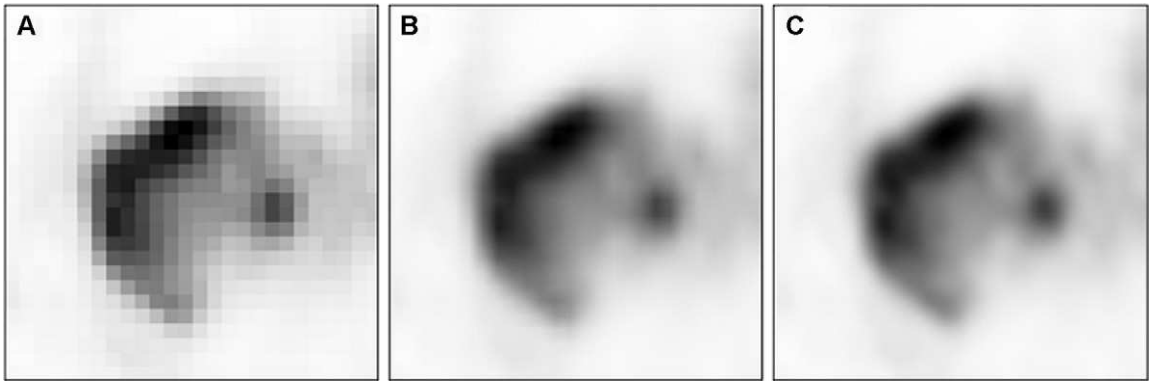



**Fig. 3.** Image de tomographie par émission de positons (TEP) d'une tumeur pulmonaire. A : originale ($5{,}31 \times 5{,}31 \times 5\,mm^3$) et interpolée sur des voxels de $1\,mm^3$ par approche B : linéaire et C : B splines cubiques.

Pour citer cet article : Hatt M, et al. Méthodologies de définition automatique des volumes métaboliquement actifs en TEP : évaluation et perspectives. Cancer Radiother (2011), doi:10.1016/j.canrad.2011.07.243

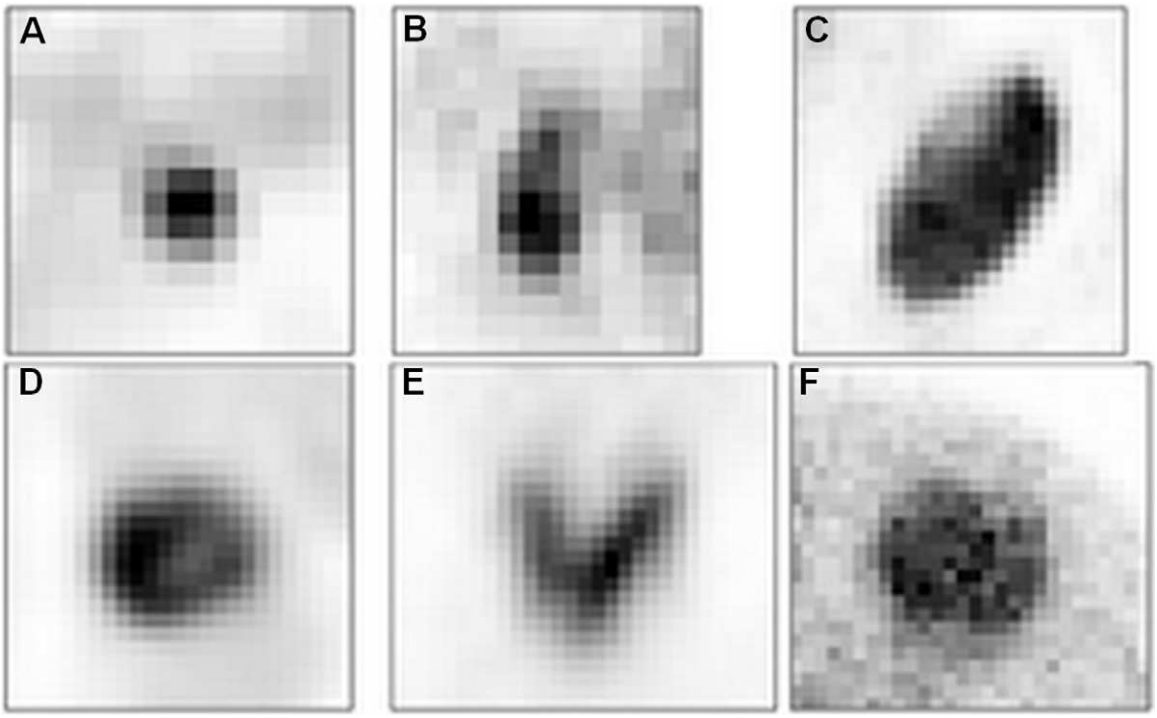



**Fig. 4.** A–C : tumeurs pulmonaires sur A : Philips Gemini ; B : Siemens Biograph et C : GE Discovery LS ; D : tumeur œsophage et E : tumeur rectale sur Philips Gemini ; et F : tumeur du sein avec ($^{18}$F)-fluorothymidine sur Siemens Biograph.

objets plus complexes en termes de formes et d'hétérogénéité, que des acquisitions de fantômes contenant des sphères. Cela peut être réalisé grâce aux simulations numériques.

Une troisième solution permettant d'évaluer les performances d'une approche est l'utilisation de données simulées réalistes. Générer de telles données nécessite l'utilisation de simulateurs de particules comme Simulation of Realistic Tridimensional Emitting Objects (SORTEO) ou Geant4 Application for Tomography Emission (GATE) associant des modèles de fantômes anthropomorphiques numériques comme Zubal ou XCAT (4D eXtended CArdiac Torso phantom) avec des modèles de systèmes d'acquisition [58–63]. Il est possible avec ces outils de générer des structures complexes et réalistes, y compris en ce qui concerne la modélisation des tumeurs [64], avec l'avantage de connaître la vérité terrain et de contrôler la plupart des paramètres. Toutefois, cette approche nécessite des compétences spécifiques et des matériels dédiés, du fait de l'exigence en termes de puissance de calcul. Des simulations analytiques plus simples peuvent également être considérées, mais les résultats doivent alors être considérés avec plus de précaution car elles impliquent des simplifications.

La mesure de mérite utilisée afin d'évaluer la précision par rapport à la vérité terrain a également son importance. Les erreurs de volume peuvent être suffisantes pour les sphères homogènes, mais ne sont pas rigoureuses pour évaluer la précision sur des objets plus complexes, car une segmentation erronée peut fournir le même volume absolu. Des mesures de mérite comme les erreurs de classification ou les coefficients de Dice sont plus pertinents [21,65].

La robustesse mesure la capacité d'une méthode à fonctionner sur la grande variabilité des images de TEP, sans ajustement préalable de paramètres. Il est donc nécessaire de considérer différents modèles de scanographe et les algorithmes de reconstruction associés, et pour chacun, différents paramètres d'acquisition pour évaluer la robustesse au bruit, au contraste, ou encore à l'échantillonnage spatial. Il est envisageable de générer de telles données avec les outils de simulations réalistes, mais cela implique la modélisation de différents scanographes et peut donc être laborieux. Une alternative consiste à exploiter des acquisitions de fantômes physiques réalisées dans différents centres, pour obtenir une large gamme de situations. Ce type de données est toutefois long à acquérir, et souffre des limites évoquées précédemment pour l'évaluation de la précision.

Évaluer la reproductibilité (aussi dénommée « répétabilité ») peut se faire par la répétition de l'exécution de la segmentation sur les mêmes données. On peut ainsi la comparer avec la variabilité observée avec la définition manuelle [18]. Ainsi, l'exécution répétée d'un seuillage fixe sur une même image donne systématiquement le même résultat avec une variabilité nulle. Au contraire, l'utilisation de méthodes avancées nécessitant par exemple un processus itératif peut mener à différents résultats en fonction, par exemple, des paramètres d'initialisation et du critère de convergence.

L'évaluation des interactions avec l'utilisateur est plus délicate à quantifier. Il est nécessaire de déterminer si l'utilisateur doit détecter et isoler le volume métaboliquement actif, définir des régions d'intérêt (dans le volume métaboliquement actif et/ou le fond), ou encore ajuster des paramètres au cas par cas. La plupart des méthodes proposées font l'hypothèse préalable que le volume métaboliquement actif a été détecté et isolé par l'utilisateur. Quelques méthodes se sont toutefois positionnées dans le cadre de la segmentation de l'image entière (Tableau 1), l'interaction avec l'utilisateur intervenant une fois la segmentation effectuée, pour analyser le résultat et associer au volume métaboliquement actif certaines des régions obtenues par la segmentation.

Par rapport à ces exigences, les méthodologies évoquées dans la section précédente fournissent des preuves de performances à des niveaux différents. Le Tableau 1 résume pour chaque approche le type de segmentation utilisée, l'interaction utilisateur, les pré- ou post-traitement nécessaires, les données utilisées pour la validation de la précision (et la vérité terrain associée), et si une évaluation de la robustesse et de la reproductibilité a été réalisée.

L'analyse de ce tableau révèle de grands écarts de qualité de validation entre les publications, ainsi qu'une grande variabilité dans les données de validation considérées. On notera le faible nombre

Pour citer cet article : Hatt M, et al. Méthodologies de définition automatique des volumes métaboliquement actifs en TEP : évaluation et perspectives. Cancer Radiother (2011), doi:10.1016/j.canrad.2011.07.243

**Tableau 1**
Comparaison des méthodologies de segmentation d'images.

| Référence(s) | Méthode | Interaction utilisateurs[a] | Pré- et post-traitement(s) | Application visée[b] | Données de validation et vérité terrain[c] | Précision sur tumeurs réalistes[d] | Évaluation de la robustesse[e] | Évaluation de la répétabilité |
|---|---|---|---|---|---|---|---|---|
| [27] | Partage des eaux | Standard + placement de plusieurs marqueurs | Ø | Globale | AF(1) - Vol. et images CT | Non | Non | Non |
| [28–31] | Gradient | Standard + initialisation | Ø | Globale | AF(5) - Vol.<br>25 TSMC - Vol. + TLG<br>18 TCH - Diam. | Oui | Oui | Oui |
| [32] | Gradient | Standard + initialisation | Débruitage et déconvolution | Globale | SF et AF - Vol. + Diam.<br>7 TCH - Complet | Oui | Non | Non |
| [33] | SA + contour actif | Standard + nécessité de fixer plusieurs paramètres | Ø | Globale | AF(1) - Vol. 1 TC - Ø | Non | Non | Oui |
| [34] | Contours actifs multimodaux (TEP/TDM) | Standard + initialisation de la forme du modèle déformable, sélection de poids | Normalisation et recalage des données de TEP et de tomodensitométrie, déconvolution des images de TEP | GTV TEP/TDM | AF - Vol.<br>2 TC - CM(1), SF | Non | Non | Non |
| [35] | Théorie des possibilités sur projections MIP | Standard | Ø | Globale | AF(1) - Vol.<br>5 TSMC – Vox.<br>7 TCH - Complet | Oui | Non | Non |
| [36] | Fuzzy C-Means (FCM) amélioré | Interprétation des classes de la segmentation effectuée sur l'image entière | Débruitage, transformations en ondelettes | Globale | 3 TSA – Vox.<br>21 TCH - Diam.<br>7 TCH - Complet | Oui | Non | Non |
| [39] | Mélange de gaussiennes | Standard + initialisation du modèle, choix du nombre de classes | Ø | Tumeurs pulmonaires | 7 TC - Ø | Non | Non | Oui |
| [40] | Champs de Markov multi-résolution | Choix du nombre de classes + interprétation des classes de la segmentation effectuée sur l'image entière | Transformations en ondelettes | Globale | AF(1) - Vol.<br>3 TC - Ø | Non | Non | Non |
| [41] | Chaînes de Markov floues (FHMC) | Standard | Ø | Globale | SF(1) et AF(2) - Vox. | Non | Non | Non |
| [18,21,22,26] | *Fuzzy locally adaptive bayesian* (FLAB) | Standard | Ø | Globale | SF (1) et AF (4) - Vox.<br>20 TSMC - Vox.<br>18 TCH - Diam. | Oui | Oui | Oui |
| [38] | Croissance de région sur critères statistiques | Standard | Optimisation sur chaque système nécessaire | Tumeurs rectales | 18 TC - CM(1) | Non | Non | Non |
| [42,43] | Arbres de décision avec apprentissage sur paramètres de texture TEP/TDM | Interprétation de la segmentation finale réalisée sur l'image entière | Apprentissage et construction des arbres de décision | GTV ORL | 10 TC - CM(3) | Non | Non | Non |
| [44] | Réseau de neurones | Interprétation des classes de la segmentation effectuée sur l'image entière | Apprentissage et construction du réseau de neurones | Globale | AF(1) - Vol.<br>3 TSA - Vox.<br>1 TCH - Diam. | Non | Non | Oui |
| [45] | Algorithme *Spherical Mean Shift* | Standard | Rééchantillonnage dans un domaine de coordonnées différent | Globale | 280 TSA - Vox. | Non | Non | Non |
| [46] | Classification des voxels à partir des courbes temps–activité | Standard + initialisation et choix du nombre de classes | Imagerie dynamique uniquement, débruitage et déconvolution | Tumeurs rectales TEP dynamique | AF - Vol. + Diam.<br>21 TC - CM(1) | Non | Non | Non |

[a] Standard : interaction « standard » (détection et placement du VMA dans une région d'intérêt).

[b] Globale : non restreint à une application.

[c] AF($x$) : acquisitions de fantôme sur $x$ scanographes différents ; SF($x$) : simulations de fantôme sur $x$ scanographes différents ; TSA : tumeurs simulées de façon analytique ; TSMC : tumeurs simulées par approche Monte Carlo ; TC : tumeurs cliniques ; TCH : tumeurs cliniques avec histopathologie ; Vol. : volume uniquement ; Vox. : voxel à voxel ; Diam : diamètre maximum uniquement ; complet : reconstruction histopathologique 3D ; CM($x$) : contours manuels ($x$ expert(s)) ; SF : seuillage fixe.

[d] Fortement hétérogènes, formes complexes, faibles contrastes, etc. avec vérité terrain rigoureuse.

[e] Requiert de multiples acquisitions sur plusieurs systèmes et un grand nombre de paramètres.

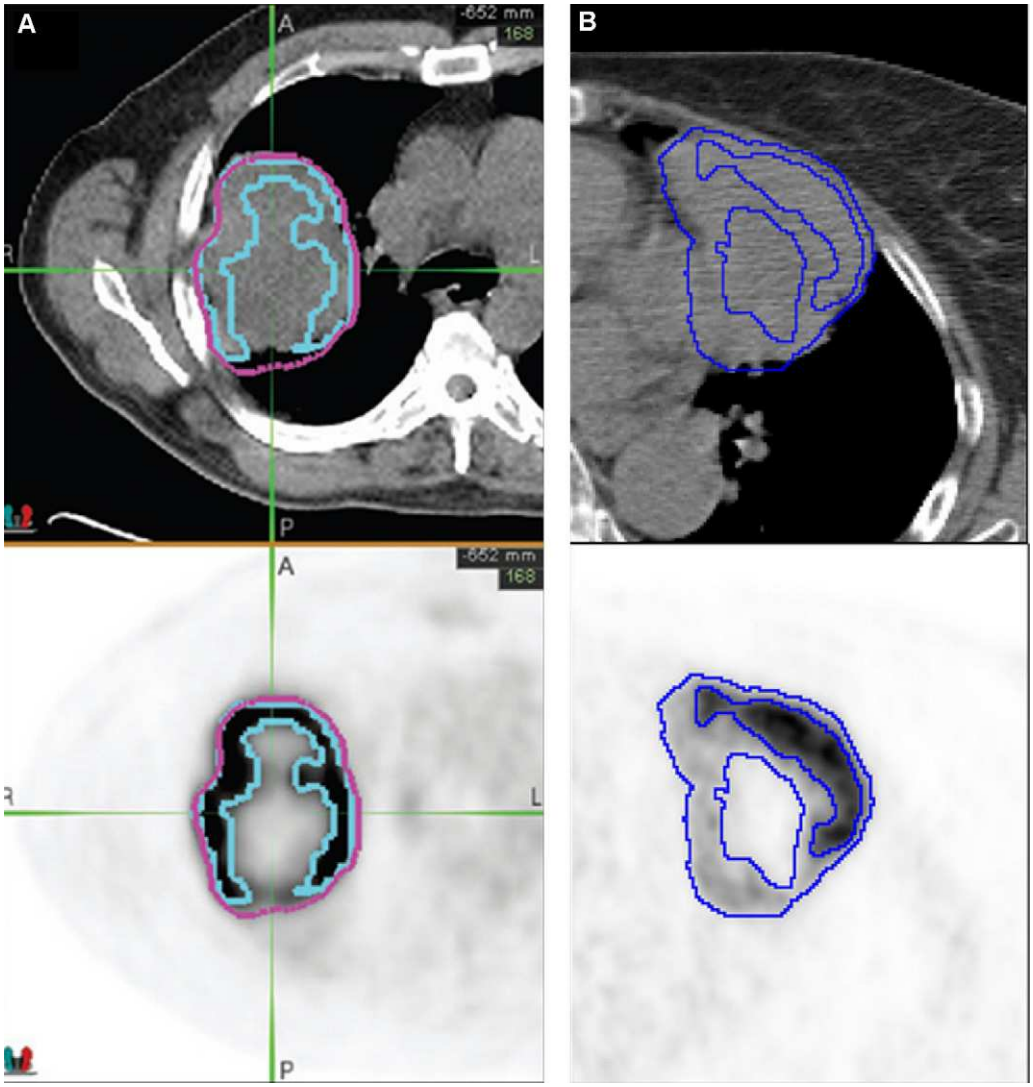



**Fig. 5.** Sur des tumeurs similaires et fortement hétérogènes, A : d'après [31] méthode par gradient (rose) qui englobe toute la tumeur sans différencier le cœur du pourtour et un seuil à 37 % (bleu) qui ne prend que le pourtour ; B : *Fuzzy locally adaptive bayesian* (FLAB) avec trois classes [26] qui définit le contour externe, le cœur nécrosé, et la fixation de plus haute activité.

de publications ayant explicitement démontré la capacité de la méthode à segmenter avec précision des volumes métaboliquement actifs réalistes dont la fixation de radiotraceur est fortement hétérogène, c'est-à-dire dont les contrastes intra-tumoraux sont suffisamment élevés pour rendre inadéquate une segmentation comme celle des seuillages [26,29,36]. Notons le cas particulier de la méthode par gradient, qui semble capable de générer un contour sur une tumeur hétérogène contenant un cœur nécrosé, mais n'est pas capable de définir le contour interne de la nécrose en question, ou de définir des régions différentes au sein de la tumeur (Fig. 5). Pour une majorité des publications, la précision a été évaluée sur des jeux de données relativement simplistes et de taille limitée et/ou des données cliniques sans vérité terrain rigoureuse. Plusieurs approches ont été appliquées sur des jeux de données disposant de mesures histopathologiques en commun. La méthode par gradient, le FCM, et celle de la théorie des possibilités ont été testées sur sept patients atteints de cancer de la sphère ORL [53], avec des erreurs par rapport aux volumes de $19 \pm 22$ %, $9 \pm 28$ % et $18 \pm 10$ % respectivement. Les méthodes FCM et FLAB ont été appliquées sur 18 patients avec tumeurs pulmonaires [14], avec $\pm 6$ % d'erreur par rapport au diamètre maximum pour FLAB [26] et $\pm 15$ % pour FCM [36] (Fig. 6). Dans la majorité des publications, les méthodes proposées ont été comparées avec des seuillages, et non avec une méthode plus performante déjà publiée. Cela explique en partie la multiplication des méthodes proposées. Il est en effet plus facile de démontrer des résultats améliorés par rapport aux seuillages, dont la précision et la robustesse sont très limitées.

La plupart des études n'ont pas apporté d'informations pertinentes sur les aspects de robustesse et de répétabilité, à l'exception des méthodes par gradient de MIMvista et FLAB. Quelques résultats concernant la répétabilité sont également disponibles dans le cas du contour actif, des réseaux de neurones et du mélange de distributions gaussiennes [33,39,44].

En conclusion, les méthodologies ayant bénéficié de la validation la plus complète (fantômes sur plusieurs scanographes pour la robustesse, images simulées réalistes, images cliniques avec histopathologie, répétabilité) sont les méthodes par gradient et FLAB.

### 3.4. *Démonstration de l'impact clinique*

Un enjeu important d'actualité est l'évaluation de l'impact clinique d'une définition précise, fiable et reproductible des volumes métaboliquement actifs pour les différentes applications de la TEP. Cela est crucial afin de convaincre les cliniciens et les industriels

Pour citer cet article : Hatt M, et al. Méthodologies de définition automatique des volumes métaboliquement actifs en TEP : évaluation et perspectives. Cancer Radiother (2011), doi:10.1016/j.canrad.2011.07.243

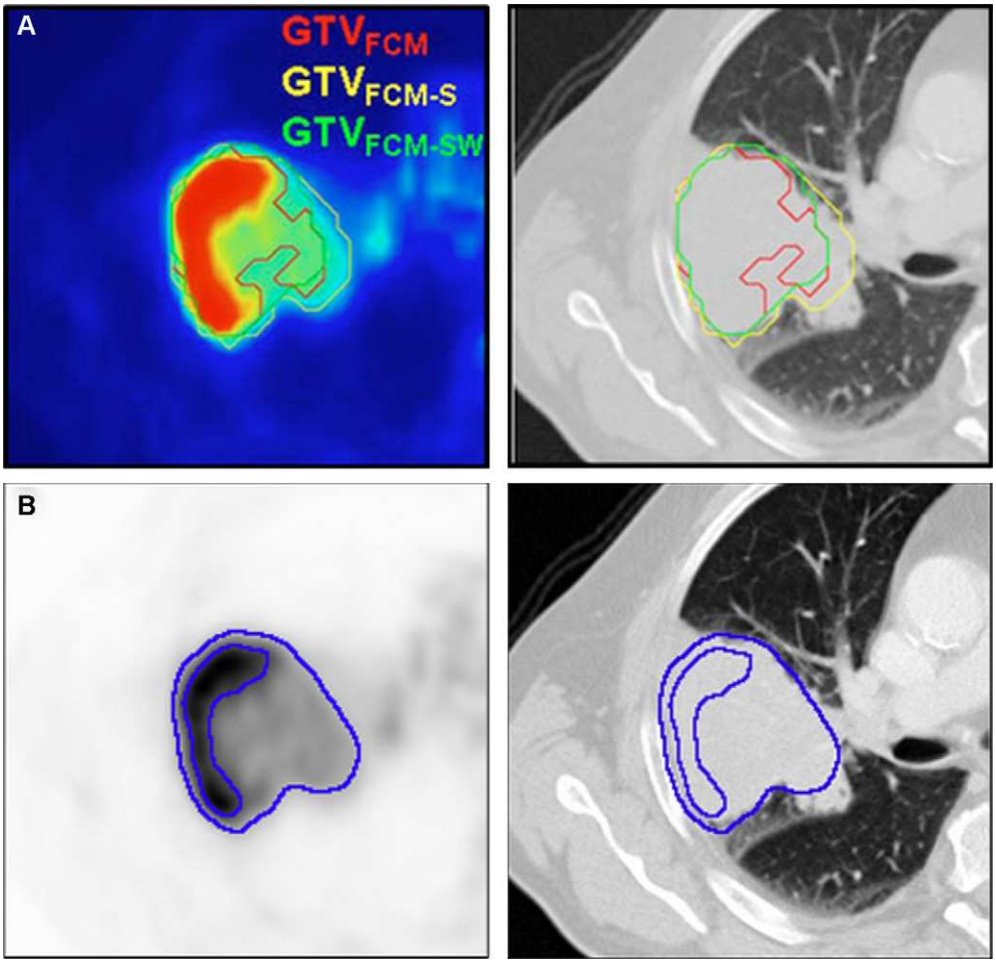



**Fig. 6.** Sur une même tumeur hétérogène, A : *Fuzzy C-Means* (FCM) amélioré [36] et B : FLAB [26]. Les contours rouge, jaune et vert correspondent respectivement au FCM simple, au FCM-S (avec corrélation spatiale), et au FCM-SW (rajoutant la gestion des hétérogénéités), qui sous-évalue la partie droite de la tumeur.

de mettre en œuvre ces approches en routine. À ce jour, seuls quelques travaux sont disponibles, démontrant, par exemple, un impact en dosimétrie pour la planification de radiothérapie utilisant la méthode par gradient [66]. Des travaux récents démontrent également l'intérêt d'une précision accrue dans la définition des volumes métaboliquement actifs, permettant d'extraire des images des paramètres tels que le volume métaboliquement actif et le *total lesion glycolysis* (TLG) associé [67]. Il a été démontré que ces derniers, contrairement aux mesures classiques de SUV, peuvent avoir une valeur prédictive de la survie et de la réponse thérapeutique dans le cadre des lymphomes, des mésothéliomes et des cancers localement évolués de l'œsophage, et ce, sur l'image prétraitement uniquement [6,9,68,69]. Ces paramètres nécessitent toutefois, contrairement à la mesure de $SUV_{max}$, une définition précise des volumes métaboliquement actifs. Notons également que cela permet d'envisager la caractérisation de l'hétérogénéité du traceur au sein du volume métaboliquement actif [8,70,71].

La détermination d'intervalles de confiance permettant de caractériser la reproductibilité des mesures de volume métaboliquement actif, afin de les utiliser pour caractériser la réponse thérapeutique [2], peut se faire sur des acquisitions répétées à quelques jours d'intervalle sans traitement. Utiliser une méthode robuste permet d'atteindre le même degré de reproductibilité que le $SUV_{max}$ ($\pm 30\,\%$), contrairement à l'utilisation de seuillages qui mènent à des niveaux de variabilité nettement plus élevés ($\pm 35$ à $\pm 94\,\%$) [18,72].

La grande majorité des méthodologies de segmentation d'images de TEP qui ont été publiées ces dernières années n'ont pour l'instant pas encore été utilisées afin de démontrer l'intérêt d'une définition fiable des volumes métaboliquement actifs dans les différentes applications cliniques, ce qui conduit à retarder leur adoption par l'industrie, et donc a fortiori par les cliniciens.

## 4. Perspectives

Comme nous l'avons exposé, la problématique de la définition automatique (ou du moins semi-automatique) des volumes métaboliquement actifs sur les images de TEP a été en partie résolue par les travaux de certains auteurs, y compris pour des situations relativement complexes de formes et d'hétérogénéité, de faibles contrastes ou rapports signal sur bruit. Les difficultés résident essentiellement dans la validation, souvent délicate et controversée, et dans l'étape de transfert à l'utilisation clinique, ce que peu de groupes ont jusqu'à présent réalisé en utilisant leurs méthodes respectives. Ajoutons à cela la popularité des seuillages fixes et adaptatifs dans les publications, qui parasitent fortement la diffusion et l'acceptation au sein de la communauté clinique de méthodologies plus performantes.

Certaines difficultés pratiques restent pour l'instant non résolues. Citons, en particulier, outre la problématique de la spécificité du radiotraceur, la différentiation automatique des fixations pathologiques et physiologiques. En ce qui concerne l'identification du volume métaboliquement actif à segmenter, il est probable que

Pour citer cet article : Hatt M, et al. Méthodologies de définition automatique des volumes métaboliquement actifs en TEP : évaluation et perspectives. Cancer Radiother (2011), doi:10.1016/j.canrad.2011.07.243

l'intervention de l'utilisateur restera nécessaire, en particulier pour des cas complexes de tumeurs situées à proximité de zones ou d'organes associés à une fixation physiologique élevée. La robustesse des méthodes face au manque actuel de standardisation des protocoles d'acquisitions est loin d'être démontrée, bien que certaines investigations aient déjà été menées [22]. Les faibles niveaux de contraste et les hauts niveaux de bruit (ou de fixations physiologiques) associés à l'utilisation de traceurs différents du FDG sont encore, même pour les approches les plus performantes, des limites complexes à dépasser. De plus, toutes les méthodes partagent des limitations techniques en termes d'initialisation de paramètres et de dépendance à l'utilisateur et peu d'entre elles ont fait la démonstration d'une automatisation suffisante pour permettre une utilisation aisée et rapide par les cliniciens. Cela est toutefois un obstacle surmontable car de nombreuses solutions d'interface utilisateur et d'estimation automatique existent, permettant de limiter les interventions de l'utilisateur ou de les rendre plus reproductibles. Il s'agit toutefois d'un effort à fournir essentiellement de la part des industriels pour la mise en œuvre des méthodologies développées par les équipes de recherche au sein de leurs produits destinés aux cliniciens.

Pour ce faire, les industriels doivent pouvoir identifier les méthodologies les plus prometteuses, ce qui est délicat, car la comparaison des méthodes est sujette à controverse si elle est réalisée sur la base des publications disponibles, faute de données de test communes. La mise à disposition de larges bases de données contenant des données cliniques associées à une vérité terrain rigoureuse comme les données d'histopathologie, et des données simulées réalistes couvrant une vaste gamme de situations, permettrait de mettre en place de telles études comparatives. Cet effort n'est pour le moment consenti que par quelques équipes regroupées au sein de collaborations limitées, bien que certaines initiatives soient déjà menées pour tenter de construire et mettre à disposition des bases de données plus conséquentes [73].

Les développements actuels de l'imagerie multimodale au sens large génèrent aussi de nouveaux défis que les méthodes développées jusqu'à présent ne permettent pas de prendre en compte explicitement. La tendance actuelle est en effet à l'augmentation du nombre de modalités d'imagerie disponibles (IRM, TEP, tomodensitométrie, etc.) et des modes d'acquisitions (radiotraceurs en TEP, séquences en IRM, etc.). Cela entraîne potentiellement la multiplication d'examens pour un patient donné, et les cliniciens sont confrontés à la prise en compte de multiples images, éventuellement associées, mais souvent acquises dans des configurations spatiales et temporelles différentes. La prise en compte automatique de ces données hétérogènes et multi-sources (plusieurs modalités et/ou plusieurs modes d'acquisition ou traceurs pour une modalité donnée, ainsi que l'évolution temporelle à différents instants d'un traitement par exemple) pour un même patient, devra donc faire l'objet de développements appropriés, pour lesquels des outils de recalage et d'analyse d'image et de données existent mais doivent être adaptés et validés.

## 5. Conclusion

Il n'existe pour l'instant pas de consensus dans la communauté sur la méthodologie à adopter pour définir automatiquement les volumes métaboliquement actifs sur les images TEP, que ce soit pour des applications de suivi thérapeutique ou la définition de nouveaux critères pronostiques et prédictifs en oncologie, ou bien la définition des volumes tumoraux macroscopiques en radiothérapie. Malgré l'existence de méthodologies ayant démontré des performances largement supérieures aux seuillages, qui restent le standard, le manque d'études approfondies et comparatives sur des données établies comme références en est la principale raison. Ce manque s'explique principalement par une ignorance des méthodologies existantes de la part d'une majorité de cliniciens, et d'une persistance de la communauté à n'utiliser et populariser les seuillages. Il s'explique également par un manque de bases de données disponibles et ouvertes, sur lesquelles chaque groupe pourrait tester les performances d'une méthode développée et la comparer à celles proposées précédemment. Tant que de telles données et études ne sont pas disponibles, il sera difficile d'obtenir des industriels une implémentation de méthodologies performantes au sein des outils destinés au cliniciens pour leur pratique routinière.

La plupart des méthodologies existantes souffrent de défauts plus au moins fondamentaux et importants, et des travaux sont encore nécessaires, notamment en termes d'automatisation et de fiabilité. Par ailleurs, de nouveaux défis voient le jour avec le développement de l'imagerie TEP multi-traceurs et les imageries multi modalité (TEP/tomodensitométrie, TEMP/tomodensitométrie, TEP/IRM, etc.), notamment pour le traitement d'informations multidimensionnelles et multi-résolution, nécessitant le développement d'approches d'analyse d'images appropriées et innovantes.

## Déclaration d'intérêts

Les auteurs déclarent ne pas avoir de conflits d'intérêts en relation avec cet article.

## Références

[1] Jerusalem G, Hustinx R, Beguin Y, Fillet G. The value of positron emission tomography (PET) imaging in disease staging and therapy assessment. Ann Oncol 2002;13:227–34.
[2] Wahl RL, Jacene H, Kasamon Y, Lodge MA. From RECIST to PERCIST: evolving considerations for PET response criteria in solid tumors. J Nucl Med 2009;50:122S–50S.
[3] Jarritt PH, Carson KJ, Hounsell AR, Visvikis D. The role of PET/CT scanning in radiotherapy planning. Br J Radiol 2006;79:S27–35.
[4] Boellaard R. Standards for PET image acquisition and quantitative data analysis. J Nucl Med 2009;50:11S–20S.
[5] Lucignani G. SUV and segmentation: pressing challenges in tumour assessment and treatment. Eur J Nucl Med Mol Imaging 2009;36:715–20.
[6] Hatt M, Visvikis D, Albarghach N, Tixier F, Pradier O, Cheze-le Rest C. Prognostic value of 18F-FDG PET image-based parameters in esophageal cancer: impact of tumor delineation methodology. Eur J Nucl Med Mol Imaging 2011;38:1191–202.
[7] Lucignani G, Larson SM. Doctor, what does my future hold? The prognostic value of FDG-PET in solid tumours. Eur J Nucl Med Mol Imaging 2010;37:1032–8.
[8] Tixier F, Le Rest CC, Hatt M, Albarghach N, Pradier O, Metges JP, et al. Intratumor heterogeneity characterized by textural features on baseline 18F-FDG PET images predicts response to concomitant radiochemotherapy in esophageal cancer. J Nucl Med 2011;52:369–78.
[9] Hatt M, Visvikis D, Pradier O, Cheze-le Rest C. Baseline (18)F-FDG PET image-derived parameters for therapy response prediction in oesophageal cancer. Eur J Nucl Med Mol Imaging 2011;38:1595–606.
[10] Pan T, Mawlawi O. PET/CT in radiation oncology. Med Phys 2008;35:4955–66.
[11] Sovik A, Malinen E, Olsen DR. Strategies for biologic image-guided dose escalation: a review. Int J Radiat Oncol Biol Phys 2009;73:650–8.
[12] Supiot S, Lisbona A, Paris F, Azria D, Fenoglietto P. *Dose-painting* : mythe or réalité ? Cancer Radiother 2010;14:554–62.
[13] Fox JL, Rengan R, O'Meara W, Yorke E, Erdi Y, Nehmeh S, et al. Does registration of PET and planning CT images decrease interobserver and intraobserver variation in delineating tumor volumes for non-small-cell lung cancer? Int J Radiat Oncol Biol Phys 2005;62:70–5.
[14] van Baardwijk A, Bosmans G, Boersma L, Buijsen J, Wanders S, Hochstenbag M, et al. PET-CT-based auto-contouring in non-small-cell lung cancer correlates with pathology and reduces interobserver variability in the delineation of the primary tumor and involved nodal volumes. Int J Radiat Oncol Biol Phys 2007;68:771–8.
[15] Ashamalla H, Rafla S, Parikh K, Mokhtar B, Goswami G, Kambam S, et al. The contribution of integrated PET/CT to the evolving definition of treatment volumes in radiation treatment planning in lung cancer. Int J Radiat Oncol Biol Phys 2005;63:1016–23.
[16] Schreurs LM, Busz DM, Paardekooper GM, Beukema JC, Jager PL, Van der Jagt EJ, et al. Impact of 18-fluorodeoxyglucose positron emission tomography on computed tomography defined target volumes in radiation treatment planning of esophageal cancer: reduction in geographic misses with equal

Pour citer cet article : Hatt M, et al. Méthodologies de définition automatique des volumes métaboliquement actifs en TEP : évaluation et perspectives. Cancer Radiother (2011), doi:10.1016/j.canrad.2011.07.243

inter-observer variability: PET/CT improves esophageal target definition. Dis Esophagus 2010;23:493–501.
[17] Gregoire V, Haustermans K, Geets X, Roels S, Lonneux M. PET-based treatment planning in radiotherapy: a new standard? J Nucl Med 2007;48:68S–77S.
[18] Hatt M, Cheze-Le Rest C, Aboagye EO, Kenny LM, Rosso L, Turkheimer FE, et al. Reproducibility of 18F-FDG and 3'-deoxy-3'-18F-fluorothymidine PET tumor volume measurements. J Nucl Med 2010;51:1368–76.
[19] Dewalle-Vignion A, Abiad AE, Betrouni N, Hossein-Foucher C, Huglo D, Vermandel M. Les méthodes de seuillage en TEP : un état de l'art. Med Nucl 2010;34:119–31.
[20] Hatt M, Visvikis D. Defining radiotherapy target volumes using 18F-fluoro-deoxy-glucose positron emission tomography/computed tomography: still a Pandora's box?: in regard to Devic et al. (Int J Radiat Oncol Biol Phys 2010). Int J Radiat Oncol Biol Phys 2010;78:1605.
[21] Hatt M, Cheze-Le Rest C, Turzo A, Roux C, Visvikis D. A fuzzy locally adaptive Bayesian segmentation approach for volume determination in PET. IEEE Trans Med Imaging 2009;28:881–93.
[22] Hatt M, Cheze-Le Rest C, Albarghach N, Pradier O, Visvikis D. PET functional volume delineation: a robustness and repeatability study. Eur J Nucl Med Mol Imaging 2011;38:663–72.
[23] Nestle U, Kremp S, Schaefer-Schuler A, Sebastian-Welsch C, Hellwig D, Rube C, et al. Comparison of different methods for delineation of 18F-FDG PET-positive tissue for target volume definition in radiotherapy of patients with non-Small cell lung cancer. J Nucl Med 2005;46:1342–8.
[24] Biehl KJ, Kong FM, Dehdashti F, Jin JY, Mutic S, El Naqa I, et al. 18F-FDG PET definition of gross tumor volume for radiotherapy of non-small cell lung cancer: is a single standardized uptake value threshold approach appropriate? J Nucl Med 2006;47:1808–12.
[25] Ollers M, Bosmans G, van Baardwijk A, Dekker A, Lambin P, Teule J, et al. The integration of PET-CT scans from different hospitals into radiotherapy treatment planning. Radiother Oncol 2008;87:142–6.
[26] Hatt M, Cheze-Le Rest C, Descourt P, Dekker A, De Ruysscher D, Oellers M, et al. Accurate automatic delineation of heterogeneous functional volumes in positron emission tomography for oncology applications. Int J Radiat Oncol Biol Phys 2010;77:301–8.
[27] Tylski P, Bonniaud G, Decenciere E, Stawiaski J, Coulot J, Lefkopoulos D, et al. 18F-FDG PET images segmentation using morphological watershed: a phantom study. IEEE Nucl Sci Sympos Conf 2006;4:2063–7.
[28] Fogh S, Karancke J, Nelson A, McCue P, Axelrod R,Werner-Wasik W, Pathologic correlation of PET-CT based auto-contouring for radiation planning in lung cancer. 13th World Conference on Lung Cancer Meeting; San Francisco USA, august 2009.
[29] Nelson AD, Brockway KD, Nelson AS, Piper JW. PET Tumor segmentation: validation of a gradient-based method using a NSCLC PET Phantom. J Nucl Med 2009;50:1659.
[30] Nelson AD, Werner-Wasik M, Choi W, Arai Y, Faulhaber PF, Ohri N, et al. PET tumor segmentation: multi-observer validation of a gradient-based method using a NSCLC PET Phantom. Int J Radiat Oncol Biol Phys 2009;75: S627.
[31] Shen G, Nelson D, Adler L. PET Tumor segmentation: comparison of gradient-based algorithm to constant threshold algorithm. Medical Physics 2007;34:2395.
[32] Geets X, Lee JA, Bol A, Lonneux M, Grégoire V. A gradient-based method for segmenting FDG-PET images: methodology and validation. Eur J Nucl Med Mol Imaging 2007;34:1427–38.
[33] Li H, Thorstad WL, Biehl KJ, Laforest R, Su Y, Shoghi KI, et al. A novel PET tumor delineation method based on adaptive region-growing and dual-front active contours. Med Phys 2008;35:3711–21.
[34] El Naqa I, Yang D, Apte A, Khullar D, Mutic S, Zheng J, et al. Concurrent multimodality image segmentation by active contours for radiotherapy treatment planning. Med Phys 2007;34:4738–49.
[35] Dewalle-Vignion AS, Betrouni N, Lopes R, Huglo D, Stute S, Vermandel M. A New method for volume segmentation of PET images, based on Possibility Theory. IEEE Trans Med Imaging 2011;30:409–23.
[36] Belhassen S, Zaidi H. A novel fuzzy C-means algorithm for unsupervised heterogeneous tumor quantification in PET. Med Phys 2010;37:1309–24.
[37] Pieczynski W. Modèles de Markov en traitement d'images. Traitement Signal 2003;20:255–78.
[38] Day E, Betler J, Parda D, Reitz B, Kirichenko A, Mohammadi S, et al. A region growing method for tumor volume segmentation on PET images for rectal and anal cancer patients. Med Phys 2009;36:4349–58.
[39] Aristophanous M, Penney BC, Martel MK, Pelizzari CA. A Gaussian mixture model for definition of lung tumor volumes in positron emission tomography. Med Phys 2007;34:4223–35.
[40] Montgomery DW, Amira A, Zaidi H. Fully automated segmentation of oncological PET volumes using a combined multiscale and statistical model. Med Phys 2007;34:722–36.
[41] Hatt M, Lamare F, Boussion N, Turzo A, Collet C, Salzenstein F, et al. Fuzzy hidden Markov chains segmentation for volume determination and quantitation in PET. Phys Med Biol 2007;52:3467–91.
[42] Yu H, Caldwell C, Mah K, Mozeg D, Coregistered FDG. PET/CT-based textural characterization of head and neck cancer for radiation treatment planning. IEEE Trans Med Imaging 2009;28:374–83.
[43] Yu H, Caldwell C, Mah K, Poon I, Balogh J, MacKenzie R, et al. Automated radiation targeting in head-and-neck cancer using region-based texture analysis of PET and CT images. Int J Radiat Oncol Biol Phys 2009;75:618–25.
[44] Sharif MS, Abbod M, Amira A, Zaidi H. Artificial neural network-based system for PET volume segmentation. Int J Biomed Imaging 2010;2010:105610.
[45] Sebastian TB, Manjeshwar RM, Akhurst TJ, Miller JV. Objective PET lesion segmentation using a spherical mean shift algorithm. Med Image Comput Comput Assist Interv 2006;9:782–9.
[46] Janssen MH, Aerts HJ, Ollers MC, Bosmans G, Lee JA, Buijsen J, et al. Tumor delineation based on time-activity curve differences assessed with dynamic fluorodeoxyglucose positron emission tomography-computed tomography in rectal cancer patients. Int J Radiat Oncol Biol Phys 2009;73:456–65.
[47] Boussion N, Cheze-Le Rest C, Hatt M, Visvikis D. Incorporation of wavelet-based denoising in iterative deconvolution for partial volume correction in whole-body PET imaging. Eur J Nucl Med Mol Imaging 2009;36:1064–75.
[48] Boussion N, Hatt M, Lamare F, Bizais Y, Turzo A, Cheze-Le Rest C, et al. A multiresolution image-based approach for correction of partial volume effects in emission tomography. Phys Med Biol 2006;51:1857–76.
[49] Aristophanous M, Yap JT, Killoran JH, Chen AB, Berbeco RI. Four-dimensional positron emission tomography: implications for dose painting of high-uptake regions. Int J Radiat Oncol Biol Phys 2011;80:900–8.
[50] Liu C, Pierce 2nd LA, Alessio AM, Kinahan PE. The impact of respiratory motion on tumor quantification and delineation in static PET/CT imaging. Phys Med Biol 2009;54:7345–62.
[51] Lamare F, Cheze-Le Rest C, Visvikis D. Le mouvement respiratoire en imagerie fonctionnelle du cancer : une revue des effets et des méthodes de correction. Traitement Signal 2006;23(numéro spécial TS et cancérologie):351–61.
[52] Warfield SK, Zou KH, Wells WM. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. IEEE Trans Med Imaging 2004;23:903–21.
[53] Daisne JF, Duprez T, Weynand B, Lonneux M, Hamoir M, Reychler H, et al. Tumor volume in pharyngolaryngeal squamous cell carcinoma: comparison at CT, MR imaging, and FDG PET and validation with surgical specimen. Radiology 2004;233:93–100.
[54] Yu J, Li X, Xing L, Mu D, Fu Z, Sun X, et al. Comparison of tumor volumes as determined by pathologic examination and FDG-PET/CT images of non-small-cell lung cancer: a pilot study. Int J Radiat Oncol Biol Phys 2009;75: 1468–74.
[55] Dahele M, Hwang D, Peressotti C, Sun L, Kusano M, Okhai S, et al. Developing a methodology for three-dimensional correlation of PET-CT images and whole-mount histopathology in non-small-cell lung cancer. Curr Oncol 2008;15: 62–9.
[56] Buijsen J, van den Bogaard J, Janssen MH, Bakers FC, Engelsman S, Ollers M, et al. FDG-PET provides the best correlation with the tumor specimen compared to MRI and CT in rectal cancer. Radiother Oncol 2011;98:270–6.
[57] Wu K, Ung YC, Hornby J, Freeman M, Hwang D, Tsao MS, et al. PET CT thresholds for radiotherapy target definition in non-small-cell lung cancer: how close are we to the pathologic findings? Int J Radiat Oncol Biol Phys 2010;77: 699–706.
[58] McLennan A, Reilhac A, Brady M. SORTEO: Monte Carlo-based simulator with list-mode capabilities. Conf Proc IEEE Eng Med Biol Soc 2009;2009: 3751–4.
[59] Jan S, Santin G, Strul D, Staelens S, Assie K, Autret D, et al. GATE: a simulation toolkit for PET and SPECT. Phys Med Biol 2004;49:4543–61.
[60] Zubal IG, Harrell CR, Smith EO, Rattner Z, Gindi G, Hoffer PB. Computerized three-dimensional segmented human anatomy. Med Phys 1994;21: 299–302.
[61] Segars WP, Sturgeon G, Mendonca S, Grimes J, Tsui BM. 4D XCAT phantom for multimodality imaging research. Med Phys 2010;37: 4902–15.
[62] Lartizien C, Kuntner C, Goertzen AL, Evans AC, Reilhac A. Validation of PET-SORTEO Monte Carlo simulations for the geometries of the MicroPET R4 and Focus 220 PET scanners. Phys Med Biol 2007;52:4845–62.
[63] Lamare F, Turzo A, Bizais Y, Le Rest CC, Visvikis D. Validation of a Monte Carlo simulation of the Philips Allegro/GEMINI PET systems using GATE. Phys Med Biol 2006;51:943–62.
[64] Le Maitre A, Segars W, Marache S, Reilhac A, Hatt M, Tomei S, et al. Incorporating patient-specific variability in the simulation of realistic whole-body 18F-FDG distributions for oncology applications. Proc IEEE 2009;9: 2026–38.
[65] Dice LR. Measures of the amount of ecologic association between species. Ecology 1945;26:297–302.
[66] Geets X, Lee JA, Castadot P, Bol A, Grégoire V. Rôle potentiel de la TEP-FDG pour la définition du volume tumoral macroscopique (GTV) des cancers des voies aérodigestives supérieures et du poumon. Cancer Radiother 2009;13: 594–9.
[67] Larson SM, Erdi Y, Akhurst T, Mazumdar M, Macapinlac HA, Finn RD, et al. Tumor treatment response based on visual and quantitative changes in global tumor glycolysis using PET-FDG imaging. The visual response score and the change in total lesion glycolysis. Clin Positron Imaging 1999;2: 159–71.
[68] Cazaentre T, Morschhauser F, Vermandel M, Betrouni N, Prangere T, Steinling M, et al. Pre-therapy 18F-FDG PET quantitative parameters help in predicting the response to radioimmunotherapy in non-Hodgkin lymphoma. Eur J Nucl Med Mol Imaging 2010;37:494–504.
[69] Lee HY, Hyun SH, Lee KS, Kim BT, Kim J, Shim YM, et al. Volume-based parameter of (18)F-FDG PET/CT in malignant pleural mesothelioma: prediction of therapeutic response and prognostic implications. Ann Surg Oncol 2010;17:2787–94.

Pour citer cet article : Hatt M, et al. Méthodologies de définition automatique des volumes métaboliquement actifs en TEP : évaluation et perspectives. Cancer Radiother (2011), doi:10.1016/j.canrad.2011.07.243

[70] El Naqa I, Grigsby P, Apte A, Kidd E, Donnelly E, Khullar D, et al. Exploring feature-based approaches in PET images for predicting cancer treatment outcomes. Pattern Recognit 2009;42:1162–71.
[71] Eary JF, O'Sullivan F, O'Sullivan J, Conrad EU. Spatial heterogeneity in sarcoma 18F-FDG uptake as a predictor of patient outcome. J Nucl Med 2008;49:1973–9.
[72] Frings V, de Langen AJ, Smit EF, van Velden FH, Hoekstra OS, van Tinteren H, et al. Repeatability of metabolically active volume measurements with 18F-FDG and 18F-FLT PET in non-small cell lung cancer. J Nucl Med 2010;51:1870–7.
[73] Tomei S, Reilhac A, Visvikis D, Boussion N, Odet C, Giammarile F, et al. OncoPET_DB: a freely distributed database of realistic simulated whole Body 18F-FDG PET images for oncology. IEEE Trans Nucl Sci 2010;57:246–55.

Pour citer cet article : Hatt M, et al. Méthodologies de définition automatique des volumes métaboliquement actifs en TEP : évaluation et perspectives. Cancer Radiother (2011), doi:10.1016/j.canrad.2011.07.243

## Autocontouring Versus Manual Contouring

**TO THE EDITOR:** We read with interest the study of Wu et al. (*1*) regarding autocontouring methodologies for target delineation in PET/CT for non–small cell lung cancer (NSCLC). Seventeen NSCLC tumors were delineated with both automated and manual approaches, using either combined PET/CT or CT and PET independently. As expected, manual contouring of PET uptake correlated better with the maximum diameter of the primary tumor than did autocontouring using a fixed threshold at 50% of maximum tumor uptake. We believe that this result is largely associated with the various shortcomings of fixed-threshold approaches, a point that needs to be clearly emphasized.

The authors have previously demonstrated that the best correlation between histopathology-derived maximum tumor diameters and image-derived ones was obtained using a 50% fixed threshold (*2*). This conclusion was reached by comparison with results obtained using other fixed-threshold values (from 20% to 55%), with a modest correlation of 0.77 and nonstatistically significant differences from the other fixed-threshold values tested. Most significantly, the use of a 50% fixed threshold led to differences larger than 1 cm in half the tumors considered. Such differences in maximum tumor diameter will most certainly lead to larger differences in the overall 3-dimensional volume. Considering similar comparisons based on 3-dimensional NSCLC tumor volumes determined by histopathology, other authors have demonstrated that an "optimal" threshold cannot be determined; considerable variability is seen (20%–42% [31% ± 11%] of the maximum), whereas CT-based volumes significantly overestimated the pathologic volume (*3*).

It is therefore important to emphasize that a fixed threshold (irrespective of its absolute value) is not an adequate methodology to delineate elevated uptake signal in PET, because of its binary, deterministic nature and lack of robustness versus varying contrast and noise conditions (*4,5*). To account for these widely documented literature findings concerning tumor target delineation incorporating PET uptake information, fixed thresholding should be avoided, and at the very least, methodologies considering target-to-background ratios such as adaptive thresholding (*5,6*) should be favored. Eventually, the wider availability of automatic segmentation approaches (*7–10*), some of which can account for the presence of heterogeneous tumor uptake (*7*), may improve the accuracy and reproducibility of adaptive thresholding (*11*) for determination of functional tumor volume.

Considering all these facts, we do agree with the authors that manual contouring should be preferred to autocontouring at a 50% threshold for functional tumor volume delineation. On the other hand, one should consider that manual delineation of PET uptake is not the ideal approach either, for multiple reasons. Most importantly, it represents a long process, particularly when it has to be performed in 3 dimensions, and it is inherently of low reproducibility (*11*).

We therefore recommend that future studies investigating this issue include the use of advanced image segmentation approaches (*4–10*), which have demonstrated improved performance in comparison to a fixed threshold and may therefore lead to alternative or complementary conclusions regarding the role of manual contouring. Irrespective of the performance of a segmentation algorithm, operator intervention will always be necessary to appropriately identify the functional uptake of interest and avoid the inclusion of non–tumor-specific uptake.



### REFERENCES


1. Wu K, Ung YC, Hwang D, et al. Autocontouring and manual contouring: which is the better method for target delineation using $^{18}$F-FDG PET/CT in non–small cell lung cancer? *J Nucl Med.* 2010;51:1517–1523.
2. Wu K, Ung YC, Hornby J, et al. PET CT thresholds for radiotherapy target definition in non-small-cell lung cancer: how close are we to the pathologic findings? *Int J Radiat Oncol Biol Phys.* 2010;77:699–706.
3. Yu J, Li X, Xing L, et al. Comparison of tumor volumes as determined by pathologic examination and FDG-PET/CT images of non–small-cell lung cancer: a pilot study. *Int J Radiat Oncol Biol Phys.* 2009;75:1468–1474.
4. Hatt M, Turzo A, Roux C, et al. A fuzzy Bayesian locally adaptive segmentation approach for volume determination in PET. *IEEE Trans Med Imaging.* 2009;28:881–893.
5. Nestle U, Kremp S, Schaefer-Schuler A, et al. Comparison of different methods for delineation of $^{18}$F-FDG PET-positive tissue for target volume definition in radiotherapy of patients with non-small cell lung cancer. *J Nucl Med.* 2005;46:1342–1348.
6. Daisne JF, Sibomana M, Bol A, et al. Tri-dimensional automatic segmentation of PET volumes based on measured source-to-background ratios: influence of reconstruction algorithms. *Radiother Oncol.* 2003;69:247–250.
7. Hatt M, Cheze le Rest C, Descourt P, et al. Accurate automatic delineation of heterogeneous functional volumes in positron emission tomography for oncology applications. *Int J Radiat Oncol Biol Phys.* 2010;77:301–308.
8. Geets X, Lee JA, Bol A, et al. A gradient-based method for segmenting FDG-PET images: methodology and validation. *Eur J Nucl Med Mol Imaging.* 2007;34:1427–1438.
9. Montgomery DWG, Amira A, Zaidi H. Fully automated segmentation of oncological PET volumes using a combined multiscale and statistical model. *Med Phys.* 2007;34:722–736.
10. Yu H, Caldwell C, Mah K, et al. Automated radiation targeting in head-and-neck cancer using region-based texture analysis of PET and CT images? *Int J Radiat Oncol Biol Phys.* 2009;75:618–625.
11. Hatt M, Cheze Le Rest C, Aboagye EO, et al. Reproducibility of $^{18}$F-FDG and $^{18}$F-FLT PET tumor volume measurements. *J Nucl Med.* 2010;51:1368–1376.


**Mathieu Hatt***
**Dimitris Visvikis**
**Catherine Cheze Le Rest**
**INSERM U650 LaTIM*
*Batîment 2bis (I3S), CHU Morvan,*
*5 Avenue Foch, Brest 29609, France*
*E-mail: hatt@univ-brest.fr*

DOI: 10.2967/jnumed.110.084897

**REPLY:** We thank Dr. Hatt and colleagues for their interest in and comments about our study of autocontouring and manual contouring for target delineation using $^{18}$F-FDG PET/CT in non–small cell lung cancer (NSCLC) (*1*). These authors are extremely accomplished in the use of PET/CT in NSCLC. We think their statement that a fixed threshold is not an adequate methodology because of its considerable variability is reasonable.

There are limited data on contouring the gross tumor volume (GTV) using PET/CT thresholds correlated with tumor size on histopathologic examination. Our study demonstrated that using the 50% fixed threshold for contouring GTV produced the best correlation between maximum tumor diameters and histopathologic findings (*2*). However, the 50% fixed threshold led to a larger difference in the diameter of GTV on PET, and CT-based volume significantly overestimated the pathologic volume. In fact, the window and level of CT also led to more differences in determining the CT-based volume (*2*). Much uncertainty exists regarding the most appropriate threshold to define a PET target volume in NSCLC radiation treatment planning. The use of a standardized uptake value (SUV) fixed-threshold intensity to define a tumor on PET may be inadequate for target volume definition and tends to underestimate target volumes (*3*). Nestle et al. (*4*) demonstrated that a GTV applying a threshold of 40% of the maximum SUV does not appear to be suitable for target volume delineation, although they used CT volume compared with PET volume because there was no available pathology correlation. For laryngeal tumors, the segmented volumes by the gradient-based method agreed with those delineated on the macroscopic specimens, whereas the threshold-based method overestimated the true volume by 68% (*5*).Yu et al. (*6*) have shown that the absolute SUV had no significant correlation with the GTV of pathology or tumor diameter.

The simplest method, which is widely used, is a visual interpretation of the PET scan and definition of contours as judged visually in cooperation with an experienced nuclear medicine physician (*7–9*). Another method using SUV is absolute SUV and regression function or source-to-background ratio. Hatt et al. (*10*) established the repeatability and reproducibility limits of several volume-related PET image–derived indices—namely tumor volume, mean SUV, total glycolytic volume, and total proliferative volume. Fixed and adaptive thresholding, fuzzy C-means, and fuzzy locally adaptive Bayesian (FLAB) methodology were considered for tumor volume delineation. The reproducibility of different quantitative parameters associated with functional volumes depends significantly on the delineation approach. State-of-the-art algorithms for functional volume segmentation use adaptive thresholding. The new 3-FLAB algorithm is able to extract the overall tumor from the background tissues and delineate variable-uptake regions within the tumors, with improved accuracy and robustness compared with adaptive threshold (tumor and background intensities) and fuzzy C-means. The gradient-based segmentation method applied to denoised and deblurred images proved to be more accurate than the source-to-background ratio method (*5*).

The different techniques to define tumor contour by $^{18}$F-FDG PET in radiotherapy planning resulted in substantially different volumes, especially in patients with inhomogeneous tumors (*4*). In our study, manual contouring was preferred to autocontouring at a 50% threshold for PET tumor volume delineation (*1*). However, manual delineation of functional volumes using PET images leads to high inter- and intraobserver variability (*11*). Furthermore, manual contouring is a long process when it has to be performed in 3 dimensions (*12*). As for the conclusion in our paper, when using autocontouring of the target in NSCLC, one should consider manual contouring of $^{18}$F-FDG PET to check for any missed disease that might be incompletely covered (*1*).

We agree with the recommendation of Hatt and colleagues that future studies investigating this issue should include a more accurate methodology, such as a segmentation algorithm. We also need to attain more data on functional volume compared with pathologic volume. Much more work must be done to resolve these issues concerning the delineation target of NSCLC using PET/CT, and we still must correlate with the gold standard—pathologic findings—whenever possible.

## REFERENCES


1. Wu K, Ung YC, Hwang D, et al. Autocontouring and manual contouring: which is the better method for target delineation using $^{18}$F-FDG PET/CT in non–small cell lung cancer? *J Nucl Med.* 2010;51:1517–1523.
2. Wu K, Ung YC, Hornby J, et al. Thresholds for radiotherapy target definition in non–small-cell lung cancer: how close are we to the pathologic findings? *Int J Radiat Oncol Biol Phys.* 2010;77:699–706.
3. Hatt M, Turzo A, Roux C, et al. A fuzzy Bayesian locally adaptive segmentation approach for volume determination in PET. *IEEE Trans Med Imaging.* 2009;28: 881–893.
4. Nestle U, Kremp S, Schaefer-Schuler A, et al. Comparison of different methods for delineation of $^{18}$F-FDG PET-positive tissue for target volume definition in radiotherapy of patients with non-small cell lung cancer. *J Nucl Med.* 2005;46: 1342–1348.
5. Geets X, Lee JA, Bol A, et al. A gradient-based method for segmenting FDG-PET images: methodology and validation. *Eur J Nucl Mol Imaging.* 2007;34:1427–1438.
6. Yu J, Li X, Xing L, et al. Comparison of tumor volumes as determined by pathologic examination and FDG-PET/CT images of non-small cell lung cancer: a pilot study. *Int J Radiat Oncol Biol Phys.* 2009;75:1468–1474.
7. Nestle U, Walter K, Schmidt S, et al. $^{18}$F-deoxyglucose positron emission tomography (FDG-PET) for the planning of radiotherapy in lung cancer: high impact in patients with atelectasis. *Int J Radiat Oncol Biol Phys.* 1999;44:593–597.
8. Kiffer JD, Berlangieri SU, Scott AM, et al. The contribution of $^{18}$F-fluoro-2-deoxy-glucose positron emission tomographic imaging to radiotherapy planning in lung cancer. *Lung Cancer.* 1998;19:167–177.
9. Nestle U, Hellwig D, Schmidt S, et al. 2-Deoxy-2-[$^{18}$F]fluoro-D-glucose positron emission tomography in target volume definition for radiotherapy of patients with non-small-cell lung cancer. *Mol Imaging Biol.* 2002;4:257–263.
10. Hatt M, Cheze Le Rest C, Aboagye EO, et al. Reproducibility of $^{18}$F-FDG and $^{18}$F-FLT PET tumor volume measurements. *J Nucl Med.* 2010;51:1368–1376.
11. Krak NC, Boellaard R, Hoekstra OS, et al. Effects of ROI definition and reconstruction method on quantitative outcome and applicability in a response monitoring trial. *Eur J Nucl Med Mol Imaging.* 2005;32:294–301.
12. Hatt M, Cheze le Rest C, Descourt P, et al. Accurate automatic delineation of heterogeneous functional volumes in positron emission tomography for oncology applications. *Int J Radiat Oncol Biol Phys.* 2010;77:301–308.



**Kailiang Wu**
**Yee C. Ung***
**Odette Cancer Centre*
*2075 Bayview Ave.*
*Toronto, Ontario M4N 3M5, Canada*
*E-mail: yee.ung@sunnybrook.ca*



DOI: 10.2967/jnumed.110.085399

Int. J. Radiation Oncology Biol. Phys., Vol. 78, No. 5, pp. 1605–1608, 2010

0360-3016/$–see front matter

# LETTERS TO THE EDITOR

## DEFINING RADIOTHERAPY TARGET VOLUMES USING $^{18}$F-FLUORO-DEOXY-GLUCOSE POSITRON EMISSION TOMOGRAPHY/COMPUTED TOMOGRAPHY: STILL A PANDORA'S BOX?: IN REGARD TO DEVIC *ET AL.* (*INT J RADIAT ONCOL BIOL PHYS* 2010)

*To the Editor:* We read with interest the article by Devic *et al.* (1) investigating the use of fixed thresholds to define non–small-cell lung carcinoma tumor positron emission tomography (PET) volumes exhibiting heterogeneous uptake. They found no correlation between the computed tomography–based and the PET-based volumes, and they associated the observed variations with intrinsic properties of PET acquisition rather than the segmentation choice. They also concluded that PET-based volumes should not be used for radiotherapy dose painting/boosting. Several studies recently dealt with similar issues considering fixed threshold to determine tumor metabolic volumes, showing large variability in the threshold values (2, 3). Other recent studies also showed the limitations of fixed thresholding and proposed more accurate and robust methods, from adaptive thresholding (4, 5) to advanced algorithms (6–8) capable in some cases of handling heterogeneous uptake frequently characterizing tumors treated with radiotherapy.

Fixed thresholds cannot reliably define functional volumes because of their deterministic and binary nature, whereas tumor uptake is variable, spatially heterogeneous, and dependent on a large number of data acquisition and image reconstruction parameters. We agree that additional studies are needed to better characterize the correlation between tracer uptake and underlying metabolism. However, irrespective of such correlation, differentiation of a PET volume from its background is an image segmentation issue that cannot be rigorously addressed using fixed threshold–based methodologies, which lead to inconsistent tumor volumes in most clinical cases (1–5), especially heterogeneous ones (1, 5, 8). In these cases and in the absence of appropriate segmentation tools, it may be more accurate (though less reproducible) to rely on manual delineation rather than a fixed threshold.

The use of inappropriate segmentation tools may lead to misleading conclusions regarding the potential of $^{18}$F-fluorodeoxyglucose–PET in guiding radiotherapy treatment planning or as a prognostic and predictive factor for therapy response (9). As new algorithms become available and clinical research applications show their potential, the medical equipment and software industry should implement them. Minimum standards and guidelines regarding functional volumes segmentation need to be developed by the different societies involved, first in clinical research and eventually in clinical practice. This is a slow process, and misleading conclusions as a result of the use of inappropriate approaches will only reduce further the process of making available new technology. We therefore suggest a more radical stance avoiding the use of any fixed threshold-based definition of PET metabolic tumor volumes in the future, especially if they are to be used for any PET image–guided therapy application.

Mathieu Hatt, Ph.D.
Dimitris Visvikis, Ph.D.
INSERM U650
National Institute of Health and Medical Research
Laboratoire de Traitement de l'Information Medicale
Brest, France

**doi:10.1016/j.ijrobp.2010.08.002**

1. Devic S, Tomic N, Faria S, *et al*. Defining radiotherapy target volumes using 18F-fluoro-deoxy-glucose positron emission tomography/computed tomography: Still a Pandora's box? *Int J Radiat Oncol Biol Phys* 2010. In Press.
2. Han D, Yu J, Yu Y, *et al*. Comparison of 18F-fluorothymidine and 18F-fluorodeoxyglucose PET/CT in delineating gross tumor volume by optimal threshold in patients with squamous cell carcinoma of thoracic esophagus. *Int J Radiat Oncol Biol Phys* 2010;76:1235–1241.
3. Wu K, Ung YC, Hornby J, *et al*. PET CT thresholds for radiotherapy target definition in non–small-cell lung cancer: How close are we to the pathologic findings? *Int J Radiat Oncol Biol Phys* 2010;77:699–706.
4. Daisne J-F, Sibomana M, Bol A, *et al*. Tri-dimensional automatic segmentation of PET volumes based on measured source-to-background ratios: Influence of reconstruction algorithms. *Radiother Oncol* 2003;69: 247–250.
5. Nestle U, Kremp S, Schaefer-Schuler A, *et al*. Comparison of different methods for delineation of 18F-FDG PET-positive tissue for target volume definition in radiotherapy of patients with non-small cell lung cancer. *J Nucl Med* 2005;46:1342–1348.
6. Geets X, Lee JA, Bol A, *et al*. A gradient-based method for segmenting FDG-PET images: Methodology and validation. *Eur J Nucl Med Mol Imaging* 2007;34:1427–1438.
7. Montgomery DWG, Amira A, Zaidi H. Fully automated segmentation of oncological PET volumes using a combined multiscale and statistical model. *Med Phys* 2007;34:722–736.
8. Hatt M, Cheze le Rest C, Descourt P, *et al*. Accurate automatic delineation of heterogeneous functional volumes in positron emission tomography for oncology applications. *Int J Radiat Oncol Biol Phys* 2010;77: 301–308.
9. Lucignani G. SUV and segmentation: Pressing challenges in tumor assessment and treatment. *Eur J Nucl Med Mol Imaging* 2009;36:715–720.

## RANDOMIZED COMPARISON OF WHOLE BRAIN RADIOTHERAPY, 20 GY IN FOUR DAILY FRACTIONS VERSUS 40 GY IN 20 TWICE-DAILY FRACTIONS, FOR BRAIN METASTASES. IN REGARD TO GRAHAM *ET AL.* (*INT J RADIAT ONCOL BIOL PHYS* 2010;77(3):648-54.)

Dr. Graham and colleagues are to be congratulated in persevering and completing their randomized study of accelerated whole-brain irradiation for brain metastases (1). The accelerated prescription of 40 Gy in 20 fractions given twice daily gained attention after Vecht *et al.* (2) reported that this prescription provided a median survival in non-operated patients of 26 weeks. This outcome was far superior to the results previously described in a number of randomized trials of palliative whole-brain irradiation, which typically used daily hypofractionated prescriptions. Our own attempt (3) to reproduce the experience of Vecht *et al.* failed. However, we are reassured to see the findings of our own randomized study with respect to survival (the same as conventionally hypofractionated radiation) and local control (improved by a factor of $\sim$2) duplicated in the report by Dr. Graham *et al.*

Nevertheless, how can a treatment that improves local control by a factor of 2 fail to have a positive impact on survival? Dr. Graham *et al.* (1) attribute this to the competing risk of death from uncontrolled extracranial disease. Although important, it is of interest that a quality-of-life study (4) that opened at our institution and competed with ours for prognostically favorable patients (age of $<$60 and Karnofsky performance scale of $\geq$70) provided those patients with a median survival of 18 weeks. By contrast, similar patients entering the accelerated trial at our institution achieved a median survival of 27 weeks (95% confidence interval, 19-35 weeks). We suspect that, in prognostically favorable patients, the provision of salvage treatments for intracranial relapse (as was done in our study at our institution) can improve overall survival in patients with controlled extracranial disease.

Where does this leave accelerated whole-brain irradiation? At the present time, we have been offering the accelerated regimen described by Vecht *et al.* (2) to patients whom I expect to be long- term survivors, in the hope of reducing the risks of long-term morbidity. The accelerated regimen provides a slightly lower equivalent dose in 2-Gy fractions ($EQD_2$) than a commonly prescribed daily regimen of 37.5 Gy in 15 fractions over 3 weeks ($EQD_2$ = 42.2 Gy for an alpha:beta ratio of 2 Gy) and published data (5) do suggest an increasing risk of measurable cognitive loss above an $EQD_2$ of 40 Gy. Of course, no one really knows what the dose per fraction sensitivity is for neurocognitive injury, whether or not the dose response is independent of clinical factors such as age, or whether alternative strategies such